THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Developing 3D novel edge detection and particle picking tools for electron tomography

Rubbiya Akram Ali

BSc (Mathematics, Statistics, Economics), MCS, MS (Computer Science)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

Institute for Molecular Bioscience

ABSTRACT

Over the past 3 billion years, photosynthetic organisms including higher plants, cyanobacteria and microalgae have evolved intricate light capturing interfaces capable of harnessing the huge energy resource of the sun (>2000x global energy demand) to produce the biomass, food and fuels which supports life on earth. As the global population expands, and with it food and fuel demand, it is becoming increasingly important to understand the structure and function of the complex and dynamic machinery of photosynthetic organisms. This is because these intricate photosynthetic systems provide invaluable blueprints for the design of next-generation solar-driven microalgae-based and bio-inspired artificial systems.

Microalgae-based systems can be located on non-arable land to produce food, fuel and high value products, in many cases using salt water. Microalgae systems have already achieved demonstration scale and the ability to produce crude oil at a price of ~$230 barrel. Through further optimization, renewable oil prices of $100 barrel could be achievable, opening the path to commercial deployment.

The first step of all biofuel production is light capture and optimizing its efficiency. However, this requires a detailed structural knowledge of photosynthetic interfaces spanning the cellular to atomic resolution range. The advances of diverse, multi-scale imaging techniques, including high-resolution single particle analysis (SPA), crystallography and electron tomography now provides a path to reveal the structure and dynamics of the photosynthetic machinery. Electron tomographic data provide unprecedented opportunities to resolve cells to molecular resolution while modern SPA, electron and X-ray crystallography and NMR can resolve the atomic resolution structure of proteins. This project is focused on bridging the gap between these techniques by developing advanced edge detection algorithms for automated tomogram segmentation to the molecular level, to facilitate molecular docking of atomic resolution protein structures and so, the development of atomic resolution atlases of the photosynthetic machinery.

Chapter 1 reviews the technology drivers for the development of solar fuel systems, the process of photosynthesis, advances in imaging and image processing technologies. The noise contamination of these images and more specifically the different types of noise reduction algorithms used to reduce its effects are reviewed. Edge detection and segmentation algorithms including manual and semi-automated segmentation, thresholding, gradient-based edge detectors, canny edge detetcor, the snake algorithm, the watershed transform, bilateral edge filter, Laplacian of Gaussian (LoG) and arbitrary Z-crossings are next reviewed to evaluate the state of the art in terms of tomogram segmentation and challenges that must be overvome to achieve molecular segmentation.

Chapter 2 describes the design and development of the automated bilateral edge (3D BLE) filter for detection of macromolecular complexes. 3D BLE was found to detect objects with edge widths as low as 2 pixels. The 3D BLE filter was shown to perform well and speed up automated segmentation significantly.

Chapter 3 describes the design, development and performance testing of the Rapid, automated 3D Z-crossing algorithm (RAZA). The advantage of RAZA over the 3D BLE filter is that it yields mathematically discrete contours that delineate each segmented object. In case of RAZA, each object has defined parameter values in terms of height, width, length, surface area and volume, which can be used as *structural fingerprints*. The 3D BLE produced non-connected edges, where as RAZA was able to deal with this problem. RAZA successfully detected the Golgi apparatus, mitochondria and mature insulin granule comparative to objects in a benchmarking test set of manually-segmented pancreatic beta cells and also resolved the contours of macromolecular objects such as ribosomes and the extrinsic domains of membrane proteins.

Chapter 4 further refines the concept of using *structural fingerprints* to automate the selection of specific classes of objects in tomograms. This provides a bridge between electron tomography, SPA, electron and X-ray crystallography and NMR on the path to the development of atomic resolution cellular atlases. Specifically, the novel RAZA particle selection ($RAZA^{PS}$) method was developed. $RAZA^{PS}$ defines each segmented object according to a set of measurable, geometric parameters and based on these, all similar particles of interest were searched, identified and quantified in an automated manner. As examples, GroEL truth set as well as mitochondria and ribosome-like macromolecules in cellular tomograms were used to perform automated quantification of organelles and complexes of interest. It was also successfully shown that using a few ribosomes like templates, thousands of molecules having similar structural characteristics could be extracted and contoured for tomographic sub-volume averaging.

In chapter 5 this work and future work on the path to detecting membrane proteins in membranes is summarized to direct the next phase of work of resolving photosystems in thylakoid membranes. The use of advanced direct electron detection systems and denoising filters in combination with $RAZA^{PS}$ now make this a tantalizing prospect which could significantly assist with the targeted engineering of the photosynthetic apparatus of microalgae and assist with the design of bio-inspired solar fuel systems of the future.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer-reviewed papers

Ali RA, Landsberg MJ, Knauth E, Morgan GP, Marsh BJ and Hankamer B. (2012) "A 3D Image Filter for Parameter-Free Segmentation of Macromolecular Structures from Electron Tomograms". *PLoS ONE* 7(3): e33697.

Conference abstracts

1. **Rubbiya A. Ali**, Michael J. Landsberg, Emily Knauth, Rosalba Rothnagel, Garry Morgan, Brad J.Marsh, Ben Hankamer. i3DBLE: An IMOD application of 3D edge detector, *Gordon Research Conference- 3D Electron Microscopy, June 2013. Colby-Sawyer College, New London, NH.* Oral and poster presentation.

2. **Rubbiya A. Ali**, Michael J. Landsberg, Emily Knauth, Rosalba Rothnagel, Garry Morgan, Brad J.Marsh, Ben Hankamer. 3D Bilateral edge filter: A step towards automated segmentation of electron tomograms*, Chemistry and Structural Biology Symposium (CASB), November 2011, The university of Queensland, Brisbane, Australia.* Poster presentation.

3. Emily Knauth, Andrew K Rindsmuth, **Rubbiya Ali,** Peter van der Heide, Anne L Sawyer, Rosalba Rothnagel, Michael J Landsberg, Brad Marsh, Ben Hankamer. Multiscale structural studies of photosynthesis and light harvesting in green algae, *Structural Biology Symposium, Novermber 2010. The university of Queensland, Brisbane, Australia.* Poster presentation.

Publications included in this thesis

The publication included in the thesis is described below.

Ali RA, Landsberg MJ, Knauth E, Morgan GP, Marsh BJ and Hankamer B. (2012) "A 3D Image Filter for Parameter-Free Segmentation of Macromolecular Structures from Electron Tomograms". *PLoS ONE* 7(3): e33697.

Full paper has been incorporated as Chapter 2 of the thesis. RAA conceived/designed the experiments, developed the algorithms, implemented the software and performed all experiments. RAA drafted the manuscript, and jointly revised it with co-authors; EK provided algal sample to capture tomogram; GPM was responsible for data acquisition; BH, MJL and BJM supervised the research.

Contributions by others to the thesis

Prof Ben Hankamer, Dr. Michael Landsberg and Dr. Brad Marsh supervised this research. Dr. Mehdi Mobli and Dr. Nick Hamilton contributed to this thesis in their role as thesis committee members.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

ACKNOWLEDGMENT

First of all, I would like to thank almighty Allah for giving me the opportunity to ponder on molecular mechanisms and appreciate His creations. I am also very thankful to Ahlul-Bayt for providing their nearness throughout this thesis. I would also like to say thanks to my father (Muhammad Akram Ali Qureshi) and mother (Anees Akhtar) for praying all days and nights for me. Especially, my father, who always encouraged me to do top-quality research. I fondly recall that during the first month of my PhD, everything looked difficult to me, I called him and told him that I wouldn't be able to complete PhD, and he said “when you look water from outside, you’ll feel cold, but once you jump into it, you get use to it, and you wont feel the temperature any more”. I would give the credit of my success to my parents.

I am really thankful to my lovely husband Ahmed Mehdi. It was him, who always encouraged me, and gave me confidence to do my studies. He helped me at home and gave me full support. Ahmed, Your love and wonderful support helped me to finish this journey. Please accept my gratitude and love for your dedication and understandings. During this thesis, I gained a lot from his knowledge.

I would like to thank also my supervisory team Dr. Ben Hankamer, Dr. Michael Landsberg and Dr. Brad Marsh for their time and support. Thank all of you for your guidance, support and advisement in the correct direction throughout my research. I have a special feeling of thanks for Brad, as he was the one who introduced me to Ben, and helped me to join IMB as a researcher. Michael always gave me positive and constructive feedback on my work, which helped me to shape up my project. Particularly, I would like to express my great appreciation for the excellent supervision of Prof. Ben Hankamer. I have never seen a man like him with so much commitment. During my PhD studies, I had really hard circumstances, but he always supported me. It was his support that I am now able to submit my thesis. His teaching capabilities are so excellent that you love doing research under his supervision. Ben! I can never forget to regard you as one of my best teachers. I have a sacred rank for you; as Imam Ali ibn-e-Abi Talib (A.S) said "**He who teaches me a word makes me his servant** ".

I would also like to extend my thanks to my thesis committee members, Dr. Mehdi Mobli and Dr. Nick Hamilton, for their encouragement and insightful comments. Special thanks to Amanda Carozzi, the IMB postgraduate co-ordinator. Amanda! Without your support and help, it is impossible to finish a degree. Whenever I had some issues during my candidature, I went to Ben and Amanda, and they both always had full support for me, and found ways to solve my problems. I also, want to say thanks to Garry Morgan, who gave me training to reconstruct tomogram. From the start of my PhD he has always been a big support for me. Unfortunately a colleague of mine, Peter

van der Heide recently passed away. I want to acknowledge him for his insight and expertise that greatly assisted the research. I will be ever grateful for his assistance, - his memory will be eternal. My special thanks to Emily Knauth, Janina Steinbeck, Rosalba Rothnagel, Dr. Ian Ross, Dr. Melanie Oey and all the group members, who have always been really helping and supportive.

I would also like to gratefully acknowledge Radosav Pantelic for his valuable discussions on this project, and Dr. Daniel Marshall for proofreading of manuscript presented in Chapter 2. I also wish to thank Erik Meijering (Biomedical Imaging Group Rotterdam, Erasmus MC, University Medical Center Rotterdam, The Netherlands) for providing the 3D Canny source code.

We gratefully acknowledge the Kühlbrandt lab (courtesy of Karen Davies, Max Planck Institute Frankfurt) for providing the data used in Chapter 3 and David Mastronarde (University of Colorado) for providing us with IMOD image processing libraries and guidance for the development of RAZA and RAZA^{PS}. We thank Dr. Christoph Gerle (Picobiology Institute, Department of Life Science, Graduate School of Life Science, University of Hyogo, Kamigori, Japan) and Christopher P. Arthur (FEI Company, Portland, Oregon, USA) for provision of the ATPase 2D crystal. We also thank Dr Yi-Wei and Jenson lab for providing cryo tomogram of *B. bacteriovorus* HD100 cells for testing RAZA.

Finally, how can I forget to acknowledge my beautiful 14 months old little son, Muhammad Raza Mehdi, who has always been a part of me during my last years of PhD. He is my love, his beautiful smiles, giggles; innocenct mischiefs never let me get tired. His one smile is enough to keep me fresh to work for many hours. I named two of my software (RAZA and RAZA^{PS}) after his name. Love you my life!

At the end, I would also like to acknowledge The University of Queensland Australia, the Institute for Molecular Bioscience for their financial support and Australian Microscopy & Microanalysis Research Facility at the Centre for Microscopy and Microanalysis, The University of Queensland.

Keywords

electron microscopy, cellular tomography, edge detection, noise suppression, image processing, microalgae, photosynthesis, biofuel, segmentation, particle picking.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080106, Image processing, 50%

ANZSRC code: 060112, Structural Biology (incl. Macromolecular modelling), 30%

ANZSRC code: 080301, Bioinformatics Software, 20%

Fields of Research (FoR) Classification

FoR code: 0801, Artificial Intelligence and Image Processing, 50%

FoR code: 0601, Biochemistry and Cell Biology 30%

FoR code: 0803, Computer Software 20%

DEDICATION

In memory of my late father,

Muhammad Akram Ali Qureshi (26.12.1942 – 23.11.2014),

For his love, endless support and encouragement. He was a self-made man, who spend his whole life in efforts for kids, but couldn't stay in this world long enough to see this accomplishment. May his soul rest in peace!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D/ 3D	Two/Three dimensions/dimensional
ADP	Adenosine di-phosphate
ATP	Adenosine tri-phosphate
ATPase	ATP synthase
BTL	Biomass-to-liquid
CAT/CT	Computerized Axial Tomography/ Computed Tomography
CCD	Charge-coupled device
CED	Coherence enhancing diffusion
$Cytb_6f$	Cytochrome b_6f complex
EED	Edge enhancing diffusion
EM	Electron microscope/Microscopy/Microscopic
ET	Cellular electron tomography/tomographic/tomogram
Fd	Ferredoxin
FNR	Ferredoxin-NADP+ oxidoreductase
HPS-FS	High pressure freezing freeze substitution
HydA	Hydrogenase
Bio-H2	Bio-hydrogen
LHC	Light harvesting antenna complex
LHCI	Light harvesting complex I
LHCII	Light harvesting complex II
NAD	Nonlinear anisotropic diffusion
NADPH	Nicotinamide adenine dinucleotide phosphate
NPQ	Non-photochemical quenching of chlorophyll fluorescence
PETC/PET	Photosynthetic electron transport chain
PC	Plactocynin
PQ	Plastoquinone
PQH2	Plastoquinol
PSI	Photosystem I
PSII	Photosystem II
SNR	Signal to noise ratio
SPA	Single particle analysis
TEM	Transmission electron microscope
NPQ	Non-Photochemical Quenching

CHAPTER 1

INTRODUCTION

Human and environmental health and global prosperity are faced with converging challenges including climate change, deforestation and desertification, reducing soil fertility and non-renewable resources, increasing pressure on arable land and fresh water, unsustainable consumption, and fuel insecurity. Collectively these issues left unchecked increase the likelihood of social, political and economic destabilization. According to the U. S. Census Bureau, every month the world's population is rising by ~6 million people and is forecast to rise to 9 billion (Figure 1.1) by 2050 (Bureau), while resources diminish. These factors point towards an unsustainable increase in demand for 70% more food (FAO 2009), 50% more water (OECD 2014), 50% more fuel (IEA 2010) by 2050. Competing against this is the need to reduce CO_2 emissions by ~50-80% by 2050, in order to avoid catastrophic consequences of climate change (IPCC 2014). These global issues are thus interlocked and their root cause is the consumption of the planet's finite and non-renewable resources.

The relatively small amount of arable land available and its optimal use for food production is another global challenge. Statistically, about 29.2% of the Earth's surface area is land and of this ~3.9% of land is arable and ~25% of the Earth's surface is non-arable (Stephens, Ross et al. 2010). Additionally, increased energy usage, places pressure on water resources in a rapidly developing world.

Energy is essential to drive our economy. Currently we use about 0.51ZJ yr^{-1} globally, of which oil provides 33.1%, coal 30.3%, gas 23.7%, hydroelectric 6.5%, nuclear 4.9% and renewables 1.6%. About 32% of this '*primary energy*' is used to generate electricity (Figure 1.2) but due to generation and transmission losses of about 38%, as end users only 20% of global energy demand is provided in the form of electricity and 80% as fuel (Stephens, Ross et al. 2013). While in recent years renewable electricity generation systems (e.g. Solar, wind and wave power have made considerable

progress) and many technologies have crossed into the main stream, the development of renewable fuel technologies is still in its infancy.

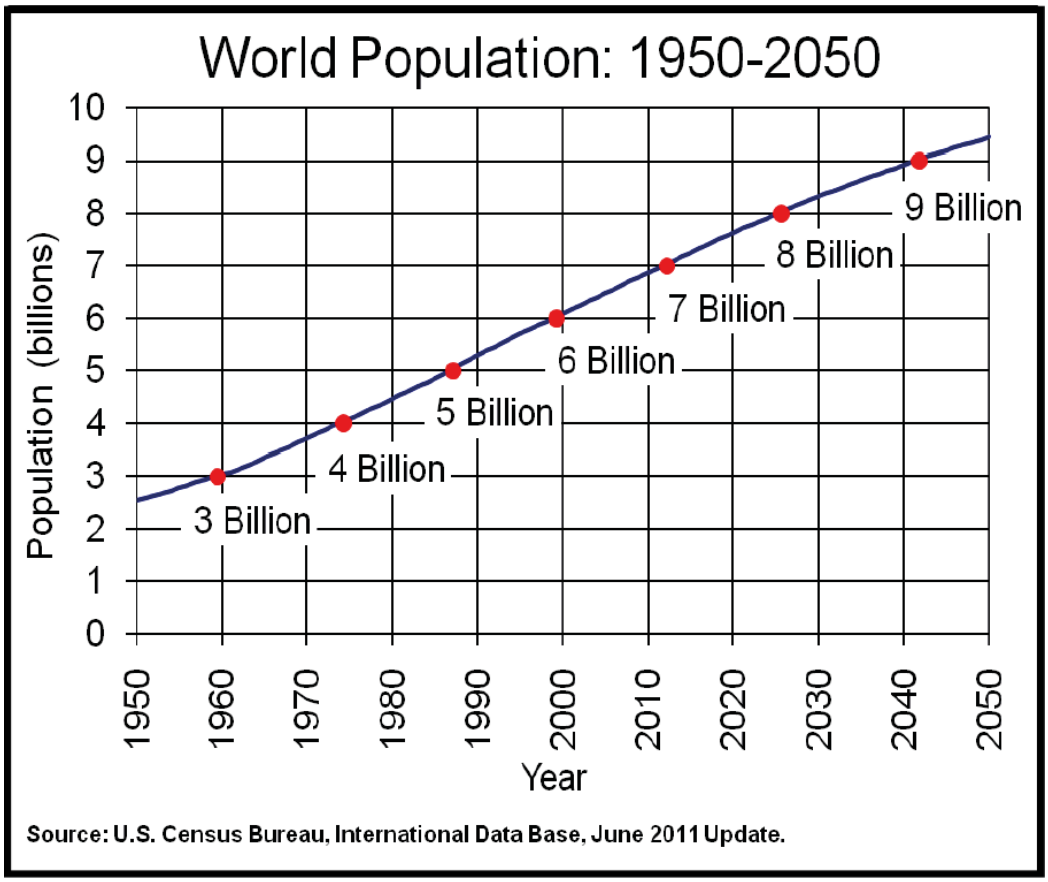


Figure 1.1 Estimated world population 1950-2050.

The production of renewable fuel on a global scale requires a sufficiently large energy source to drive its production. Solar energy is by far the largest renewable energy source available to us with 5500 ZJ yr^{-1} (1 ZJ = 10^{15} J) arriving at the Earth's atmosphere (Smil 2006). Of this 1300 ZJ yr^{-1} (or 2600 x global energy demand) reaches the Earth's surface in the form of photosynthetically active radiation. Photosynthetic cyanobacteria, microalgae and higher plants have over about 3 billion years evolved intricate and dynamic photosynthetic systems which can tap into this huge solar energy resource and use it to produce the biomass, food, fuel and oxygen which powers the biosphere and supports oxygenic life on Earth. Over the past decade, microalgae and artificial solar fuel systems have emerged as two renewable fuel front runner technologies capable of harnessing the sun's energy for the production of solar fuels. Microalgae systems are widely considered the most advanced, with pilot and demonstration systems under test internationally. This is critically important as the next 20 years requires a major reduction in CO_2 emissions if we are to remain within a 2° C global warming limit (Stocker and IPCC). Artificial solar fuel systems are inspired by the natural photosynthetic machinery and offer the potential of higher efficiencies, but are largely at

the stage of component research and development. These systems, for example, aim to couple water splitting catalysts (Brimblecombe, Dismukes et al. 2009, Dinca, Surendranath et al. 2010), light driven redox active centers (e.g. based on Photosystem II) and hydrogenases (Berggren, Adamska et al. 2013) to drive solar H_2 production from water. Using different component sets, other carbon-based fuels could also be produced.

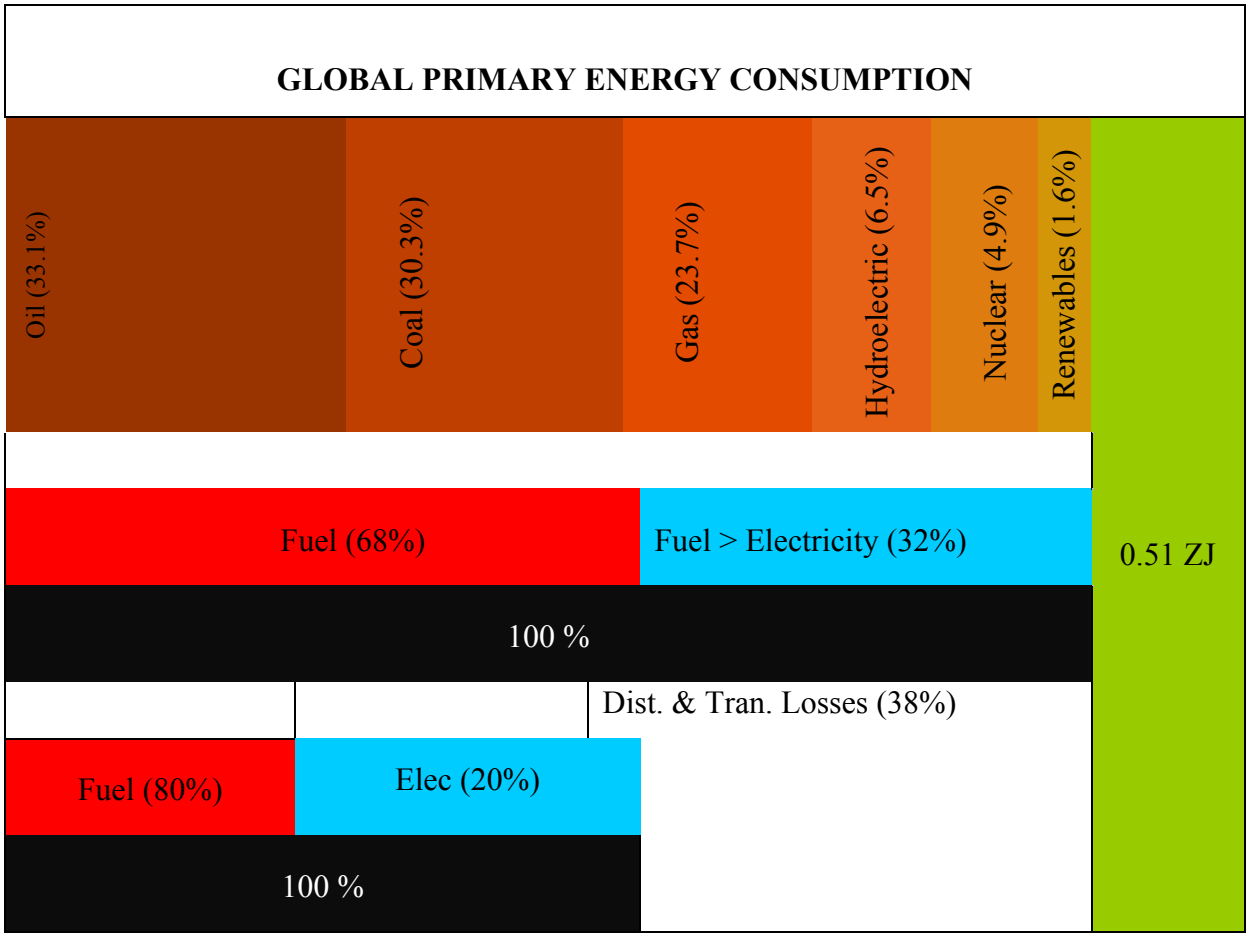


Figure 1.2 Global primary energy consumption.

1.1 Microalgal biofuels

Microalgae are now considered as one of the most promising sources for biofuel feedstock (Greenwell, Laurens et al. 2010, Hunter 2010, Wijffels and Barbosa 2010). Discussion on fast-growing microalgae (eukaryotic or cyanobacteria) as a source of bio-fuel systems started over 50 years ago. But the impending pressures of the depletion of oil, food, water, and soil fertility now provide a more concerted focus to change (Wijffels and Barbosa 2010). The first step of all microalgal biofuel-producing processes is photosynthesis, which absorbs sunlight, CO_2, H_2O and nutrients, and ultimately produces a wide range of fuel feed stocks (Hankamer, Lehr et al. 2007). Microalgal biofuel systems have the advantage that:

1) Their production is independent of soil fertility, allowing them to be produced on non-arable land.
2) They can be produced in closed or recycling systems, which greatly reduce water and nutrient inputs compared to crop based biofuels such as corn ethanol.
3) In some cases, saline water can be used to conserve valuable fresh water resources.
4) They provide economic benefits such as assisting with regional development by utilizing infertile land.
5) They can produce a wide range of renewable fuels and fuel feed stocks including biocrude oil, biodiesel, methane, butanol and hydrogen.
6) They can produce a range of high value products, which improve system economics including animal feeds and therapeutic proteins such as vaccines.
7) Microalgae can produce Agrichar^{TM}, a form of charcoal, which can be used to improve soil quality and contribute to carbon storage.

Despite the importance of the photosynthetic machinery in these processes and significant advances in the structure determination of its constituent membrane protein complexes (Stroebel, Choquet et al. 2003, Standfuss, Terwisscha van Scheltinga et al. 2005, Amunts, Toporik et al. 2010, Drop, Webber-Birungi et al. 2014, Suga, Akita et al. 2015), the organization and dynamic adjustment of the photosynthetic machinery within thylakoid membranes in response to ever changing environmental conditions is still at the early stages of elucidation. And yet, understanding these structures and processes is not only central to photosynthesis research in general, but to revealing the structural blue-print for the targeted engineering of high efficiency cell lines for biotechnological applications as well as gaining design principles for artificial solar fuel systems. The emergence of diverse, multi-scale imaging techniques and image processing tools such as high-resolution cryo-electron microscopy, single particle analysis (SPA), crystallographic and Electron tomography (ET) data sets, to reveal the structure and dynamics of protein complexes within the thylakoid membranes, offers unprecedented opportunities to build an accurate 3D atlas of the photosynthetic machinery and its change over time (4D). My project is focused on initiating the development of such a 3D atlas of the model green algae *Chlamydomonas reinhardtii (C.reinhardtii)* to ultimately enable targeted engineering of microalgal biosystems with improved light capture efficiency, and the structure-guided design of artificial solar fuel systems. To achieve this, the supramolecular structure and function of the chloroplast and the embedded photosynthetic machinery of thylakoid membranes must be resolved and modelled within a cellular context. In the next section the structure of *C. reinhardtii* and the photosynthetic machinery is summarized.

1.2 Structure of C. reinhardtii

C. reinhardtii is an oval-shaped unicellular green alga that is approximately 10 µm in diameter and 3 µm in width (Rochaix 2001). In 1939, it was reported for the first time that unicellular green algae could produce light-driven hydrogen production, highlighting early interest in solar fuels (Gaffron 1939).

Figure 1.3 shows that *C. reinhardtii* cells contain structures typical of eukaryotes including a nucleus and organelles including the mitochondria and Golgi apparatus as well as a cell membrane and cell wall. It has evolved two flagella and an eyespot which helps orient the cell to provide optimised light conditions, a large single cup-shaped chloroplast and thylakoid structures to capture light as well as a large pyrenoid and associated starch granules. The single chloroplast of *C. reinhardtii* is an intracellular organelle with a U-shaped structure. Each of these components of chloroplast is summarized in more detail below.

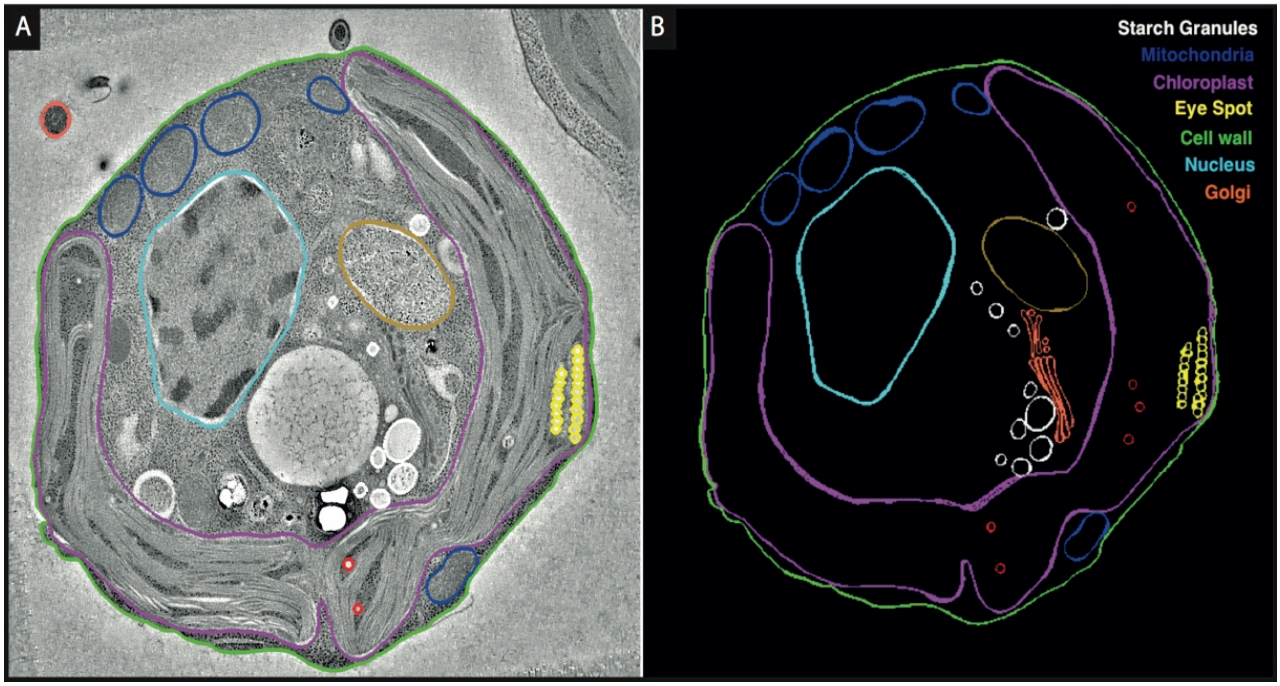


Figure 1.3 Contour modelling and volume rendering of the algal cell.

A. shows the cross section of a C. reinhardtii segmented algal cell showing compartmentalization. B. 3D surface view of the segmented algal cell. The colors outline the structures: Cell wall (green), chloroplast (purple), eye spot (yellow), nucleus (light blue) shows, dark blue shows mitochondria, white shows starch granules, orange shows Golgi.

1.2.1 The Chloroplast

The chloroplast is the organelle, which contains the photosynthetic machinery and converts sunlight into chemical energy. Details of its ultrastructure (Figure 1.4) are summarized below:

1.2.1.1 Eyespot or stigma

The eyespot is reported to be located in the chloroplast, between the front and back regions of the cell (Sager and Palade 1957). It lies directly below the chloroplast envelope and contains stacks of thick granules closely crowded together. It consists of two or more layers/plates of regularly arranged electron-dense granules, approximately 80-130 nm in diameter. A thylakoid membrane subtends each layer. Each plate is composed of a single layer of dense, spherical, uniform bodies about 100 to 140 mμ in diameter. Being embedded in the chloroplast matrix it has a definite relationship with the envelope. One plate of granules sits just beneath the chloroplast envelope, followed by a disc, a second plate of granules, and another disc. Each granule consists of a dense core and is separated by narrow spaces.

The eyespot is a light sensitive organelle, which senses the direction and intensity of light to swim. It mediates light perception, which helps the cells to find optimal photosynthetic light conditions. In *Chlamydomonas* cells, the eyespot proteome (Schmidt, Gessner et al. 2006) contains a large number (roughly 200 different proteins) of structural, signalling and metabolic proteins, for example, to sense light, it has photoreceptor proteins (Hegemann 1997, Kreimer 2009).

1.2.1.2 Pyrenoid

The pyrenoid is a differentiated region of the chloroplast stroma of many eukaryotic algae composed of micro-compartments involved in the operation of carbon-concentrating mechanism (Giordano, Beardall et al. 2005). It is a spherical body with a diameter of 1.2 to 2 µm. It appears in thin segments that are homogeneously dense. Due to it's close association with starch granules, it is believed that the pyrenoid has a function in starch synthesis (Sager and Palade 1957). It is considered to be a major component involved in the operation of a carbon-concentrating mechanism (CCM) that generates and maintains a CO_2 rich environment around the photosynthetic enzyme ribulose-1, 5-bisphosphate carboxylase/oxygenase (Rubisco) (Mukherjee and Moroney), which is localised to the pyrenoid matrix within the chloroplast in a number of green algae including *C. reinhardtii* (Morita, Kuroiwa et al. 1997).

1.2.1.3 Chloroplast envelope

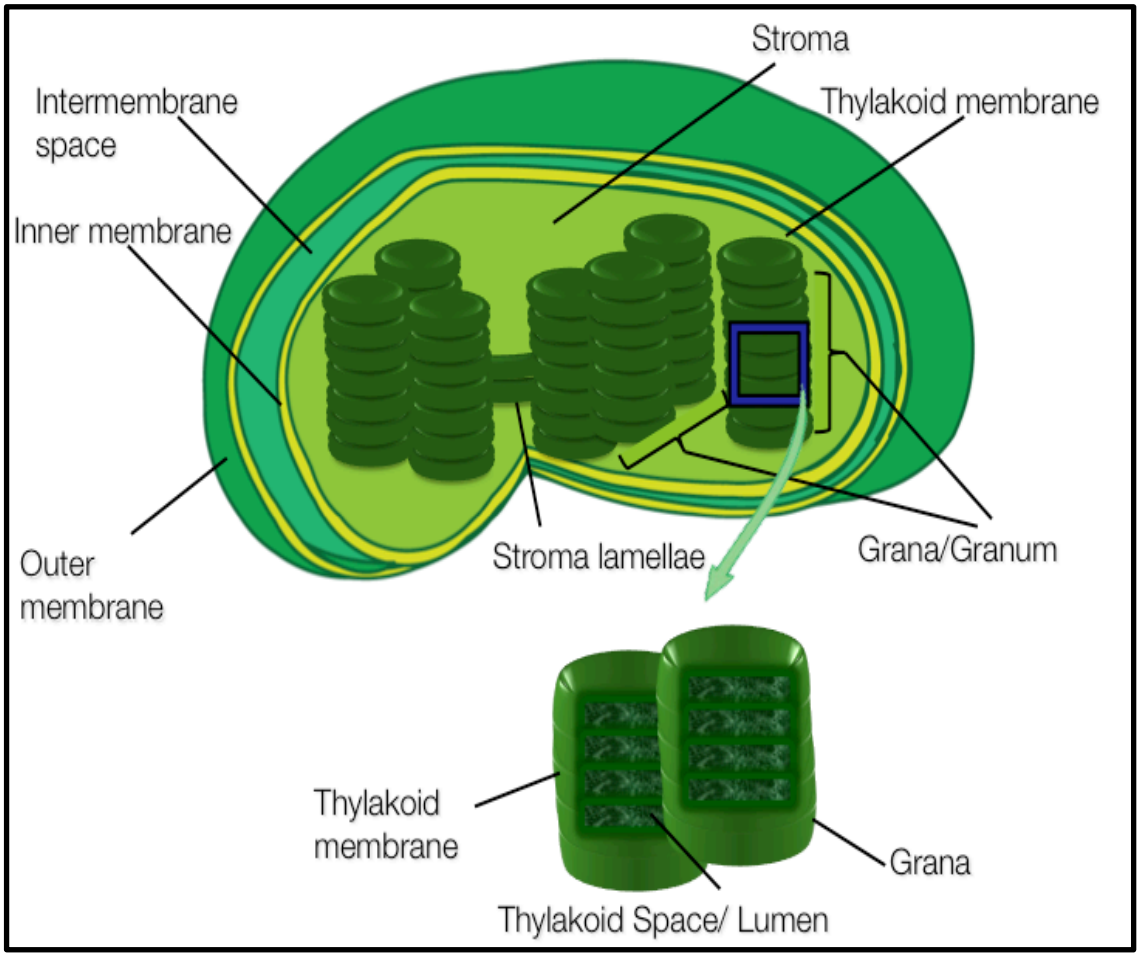
The chloroplast is protected by a double membrane structure called the chloroplast envelope. It is the interface between organelles and other cellular processes, which keeps chloroplast structures enclosed. It consists of the **inner** and **outer envelope membranes,** each ~50 Å in thickness (Sager and Palade 1957). The inner envelope membrane regulates the flux of macromolecules into and out of the chloroplast, while the outer envelope allows the transfer of low molecular weight molecules

(Heldt and Sauer 1971, Inoue 2007). The envelope not only regulates the transfer of macromolecules between the stroma and the cystol but also mediates a regulated nuclear import of proteins. Each of these membranes is a lipid bilayer with 6-8 nm thicknesses. In spinach chloroplasts, the lipid composition of the outer membrane has been found to be 48% phospholipids, 46% galactolipids and 6% sulfolipids, while the inner membrane has been found to contain 16% phospholipids, 79% galactolipids and 5% sulfolipids (Block, Dorne et al. 1983). The outer membrane is permeable to most ions and metabolites, but the inner membrane is highly specialized, containing a number of transport proteins (Heldt and Sauer 1971, Inoue 2007). For example, a triose phosphate translocator transports carbohydrates across the inner envelope membrane (Walters, Ibrahim et al. 2004). These two membranes are separated by the **inter-membrane space** with a gap of 10-20nm between them.

1.2.1.4 Thylakoid

The **stroma** is a dense fluid within the chloroplast envelope, which surrounds the thylakoid membranes and is the site where CO_2 is converted to carbohydrates (sugar). The **thylakoid membranes** are organized into granal stacks and stromal lamellae. Photosystems I and II, their corresponding light harvesting complexes (LHCI and II) as well as $cytb_6f$ and ATP synthase which are the key components of the photosynthetic electron transport chain are embedded in the thylakoid membranes (Harris 2009). The photosynthetic electron transport chain (PET) produces NADPH and ATP required for CO_2 fixation and other cellular processes.

The lipid composition of the thylakoid membrane is quite similar to the inner envelope membrane, containing 78% galactolipids, 15.5% phospholipids and 6.5% sulfolipids in spinach chloroplasts (Daum, Nicastro et al. 2010). The thylakoid membrane encloses a single, continuous aqueous compartment called the thylakoid lumen (Mustardy and Garab 2003). **Grana** are densely layered stacks of thylakoid membranes often disc like in arrangement. Typically 10 to 100 grana stacks are reported to be observed in a chloroplast, but with the advance of electron tomography this can now be analyzed in more detail. The grana are connected to each other by the **stroma lamellae.** Grana thylakoids have a different composition of proteins as compared to stroma thylakoids. Active PSII and LHCII are considered to be enriched in the grana and PSI and LHCI in the stroma. ATPases and PSII have been reported in both spinach and pea(Daum, Nicastro et al. 2010) .



1.2.1.5 Chloroplast ribosome

Ribosomes are small machines that produce proteins. Ribosomes synthesise proteins from mRNA. They do so by using the genetic information in the mRNA to translate mRNA into proteins. There

are two types ribosomes; free and membrane-bound. Free ribosomes move anywhere in the cytosol, whereas membrane bound ribosomes are bound to some organelles where there is a need to synthesize proteins. There are lots of different ribosomes structures – small and large subunits, or the entire ribosomes, while eukaryotic and prokaryotic ribosomes differ very much in their size and subunit composition.

George Palade discovered ribosomes first time in 1950s using an electron microscope. He (together with Christian de Duve and Albert Claude) was then laureate with the Noble prize in Physiology or Medicine (Boyer and Walker 1997). The prokaryotes (e.g. bacterial) (70S) ribosome consists of two subunits; 1) small (30S) and 2) large (50S). The eukaryotes have (80S) ribosomes. Each of these ribosomes also has two subunits; small (40S) and large (60S) subunits. The first structure of the ribosomes was determined in late 2000. Ben et al reported the complete atomic structure of large (50S) subunit from the *Haloarcula marismortui* (Ban, Nissen et al. 2000) using crystallography and achieved 2.4Å resolution. Similarly, the structure of ribosome from *Thermus thermophiles* was determined in 2000 using X-ray crystallography to reveal the structure of the small subunit (30S) at 3 Å resolution (Wimberly, Brodersen et al. 2000). In 2009, Venkatraman together with Thomas and Ada, determined the detail structure and function of the ribosomes, and were awarded the Noble prize in Chemistry (Ramakrishnan , Steitz et al. 2009). Most recently, Fischer et al. resolved the structure of the E. coli ribosome in complex with the elongation factor EF-Tu to 2.65Å resolution (Fischer, Neumann et al. 2015) using single particle cryo-EM.

1.3 Photosynthesis

The process of photosynthesis is fundamental to all life on earth, as it captures solar energy and stores it as chemical energy in the form of ATP and NADPH. The process commences with the absorption of photons by arrays of chlorophyll molecules (Chl a and Chl b) intricately coordinated within light harvesting complex (LHCI and LHCII) proteins and the photosystems that they connect to. The LHC proteins belong to a large gene family, which in the green alga *C. reinhardtii,* consists of over 20 members (Dittami, Michel et al. 2010). These LHC proteins perform two main tasks: To capture photons and channel the derived excitation energy to the photosynthetic reaction centers of photosystems (PSI & PSII). The excitation energy is used by PSII to split water (H_2O) into protons (H^+), electrons and oxygen (O_2). Electrons then enter the photosynthetic electron transport chain (PET - linear- Figure 1.5A *thin solid black arrows*). In this process, Plastoquinol (PQ) carries electrons from PSII to $Cytb_6f$. Plactocynin (PC) receives the electrons from $Cytb_6f$ and shuttles them to PSI. Electrons are then passed through ferredoxin (Fd) and on to NADPH.

To reduce photo damage to PSII, LHCs dissipate excess energy via the processes of Non-Photochemical Quenching (NPQ). Under high light conditions, NPQ can cause energy losses of up to ~90% (Polle, Kanakagiri et al. 2003, Mitra and Melis 2008) of the captured solar energy. For the development of efficient microalgae systems, it is important to minimize these losses and to achieve this requires a detailed understanding of the organization of individual PSII-LHCII and PSI-LHCI super complexes to define their functional importance in terms of light capture and energy dissipation.

PSII and the PQ/PQH_2 cycle simultaneously release protons into the thylakoid lumen, thereby generating a proton gradient. This proton gradient drives the production of ATP via the ATP synthase. Upon transfer to the stroma, the protons and electrons are used to reduce NADP to NADPH.

ATP and NADPH are used in the Calvin cycle and other processes to drive the production of biomass (*for bio-methane & biomass-to-liquid (BTL)*), oils (*for bio-diesel & aviation fuel*), sugars & starch (*for bio-ethanol*) and protons & electrons (*for bio-H_2* in the case of green algae (*C. reinhardtii*)). H_2 is considered to be the most efficient form of biofuel production as it is directly coupled to the photosynthetic electron transport chain, and unlike the production of carbohydrates and oils for other biofuels, requires no additional ATP or NADPH. Although H_2 is often considered to be a 'fuel for the future' it is already used extensively in industry and for synthetic fuel production (Gupta, Kumari et al. 2013). Photosynthesis in microalgae is constantly adapting to changing environmental conditions to maximize efficiency and minimize photodamage. To achieve this it undergoes a set of dynamic changes, which include the state transitions and cyclic electron transport.

1.3.1 State 1/ State 2 transition

Under low NPQ/linear electron transport conditions, the highest efficiency of ATP and NADPH is generally achieved, as all the captured energy in the form of photons is used to derive chemical energy from sunlight and only two photons drive the transfer of an electron from H_2O to NADPH. In this scenario one photon is used to excite PSII and the other one is used to excite PSI to facilitate the transfer of one electron from water to NADPH.

Microalgae and higher plants have also evolved the State 1 (Figure 1.5A) to State 2 (Figure 1.5B) transition process to maintain linear electron transport under varying light limiting conditions. Under high light conditions, the PQ pool can become over-reduced (PQH_2) if the turnover rate of PSII exceeds that of PSI and this can result in photodamage. In the case of *C. reinhardtii*, to protect against such photodamage a subset of the LHCII proteins can dissociate from PSII and migrate to PSI (Takahashi, Iwai et al. 2006, Tokutsu, Iwai et al. 2009). This transition yields a PSI-LHCI-

LHCII supercomplex (Figure 1.5B), which allows PSI to capture more light (Kargul, Turkina et al. 2005, Drop, Webber-Birungi et al. 2011) balancing the turnover of both PSI and PSII, reoxidizing the over-reduced PQH_2 pool to PQ and maintaining efficient linear electron transport between them. However, if the PQ pool continues to be over-reduced (Figure 1.5C), or the cell has a greater requirement for ATP, NADPH-dependent cyclic electron transport is reportedly induced (Murata 1969).

1.3.2 Cyclic electron transport (CET)

In CET, PSII and the PQ/PQH_2 cycle simultaneously take protons (H^+) from the stroma and release them into the thylakoid lumen, thereby generating a proton gradient. This proton gradient increases the proton gradient required to drive the production of ATP via ATP synthase. Similarly the electrons (e^-) are taken from the stromal side of PSI and cycled back via $Cytb_6f$ to PSI rather than donating them directly to NADPH, thereby reducing the rate of NADPH production and maintaining electrochemical balance across the thylakoid membrane. In so doing, CET helps to adjust the ATP/NADPH ratio to match the needs of the cell and the over reduced PQH_2 is re-oxidized to PQ to protect PSII from damage leading to photo-inhibition.

It has recently been postulated that a 1 MDa protein 'CEF-PSI supercomplex' acts as the site of cyclic electron transport in *C. reinhardtii* (Iwai, Yokono et al. 2010). This super complex (Figure 1.5C- *CEF-PSI in black box*) reportedly consists of the PSI-LHCI and $Cytb_6f$ complexes, Ferredoxin-NADPH Oxido-reductase (FNR), and the integral membrane protein PGRL1. Independent studies have revealed that PGRL1 is also required for efficient cyclic electron transfer in *C. reinhardtii* (Petroutsos, Terauchi et al. 2009, Tolleter, Ghysels et al. 2011).

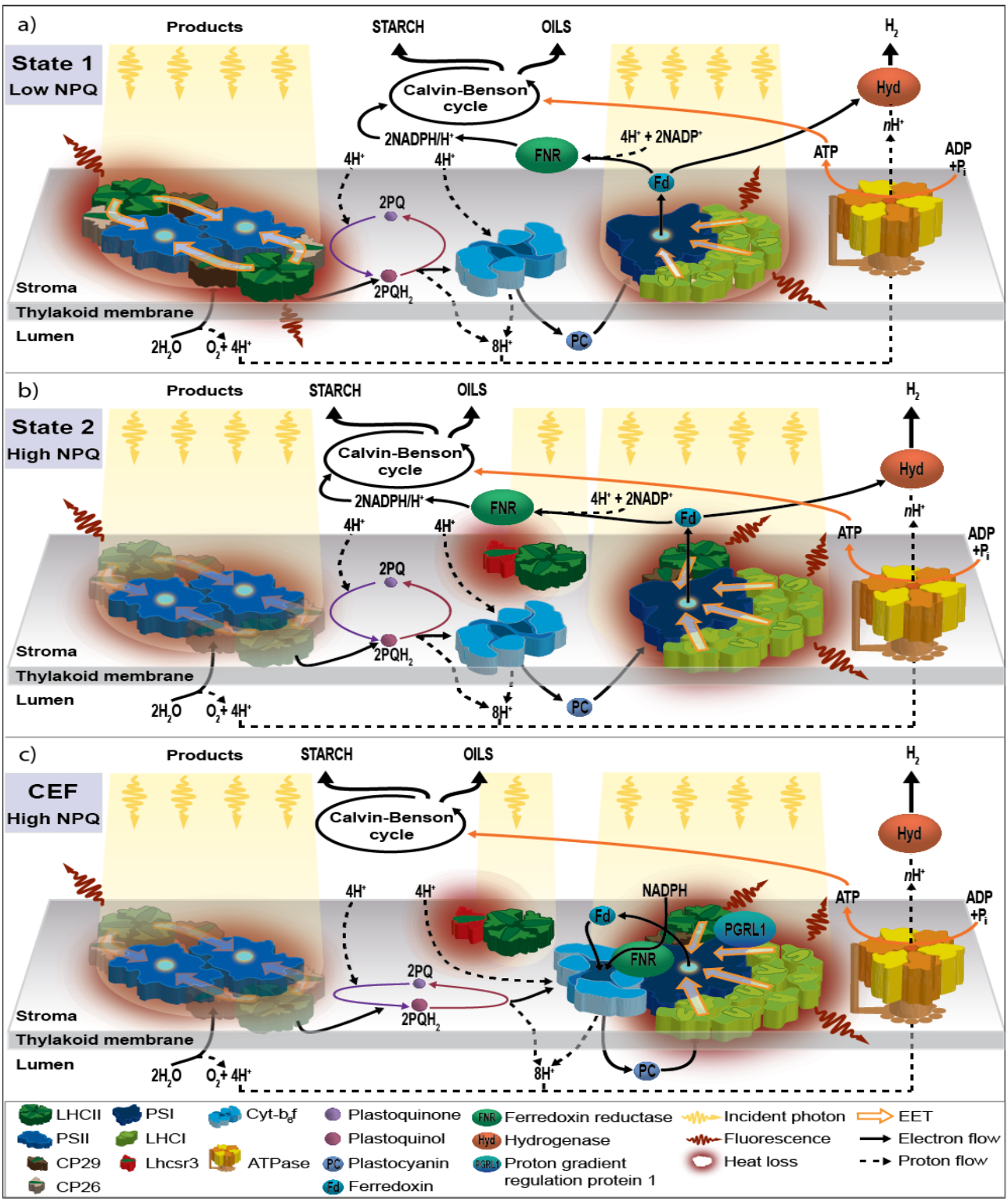


Figure 1.5 Photosynthesis process in microalgae.

A) State 1 with low NPQ. B) State1/state 2 transition with high NPQ C) Cyclic electron flow chain with high NPQ. Image courtesy of Dr. Drew Ringsmuth. used with permission.

1.3.3 Molecular organization of super complexes of photosynthetic machinery embedded in thylakoid membranes

In recent years, extensive progress has been made on solving the structures of the photosynthetic machinery using a combination of x-ray crystallography and single particle analysis. Atomic

structures are available for all key PSII, $Cytb_6f$, PSI and ATP synthase complexes, but not all are from the green alga *C. reinhardtii*. Nor are atomic structures available for the various conformational states. A brief summary of the available structural data is provided below.

1.3.3.1 Photosystem II

Photosystem II (PSII) is a multisubunit pigment-protein complex containing approximately 20 subunits including PsbA-PsbX, Lhcb1-Lhcb6) in higher plants and green algae(Figure 1.6) It is located within the thylakoid membrane (predominantly in the grana) of plants, algae and cyanobacteria, and drives the photosynthetic water splitting reaction (Renger and Renger 2008), which is the most oxidising reaction in biology.

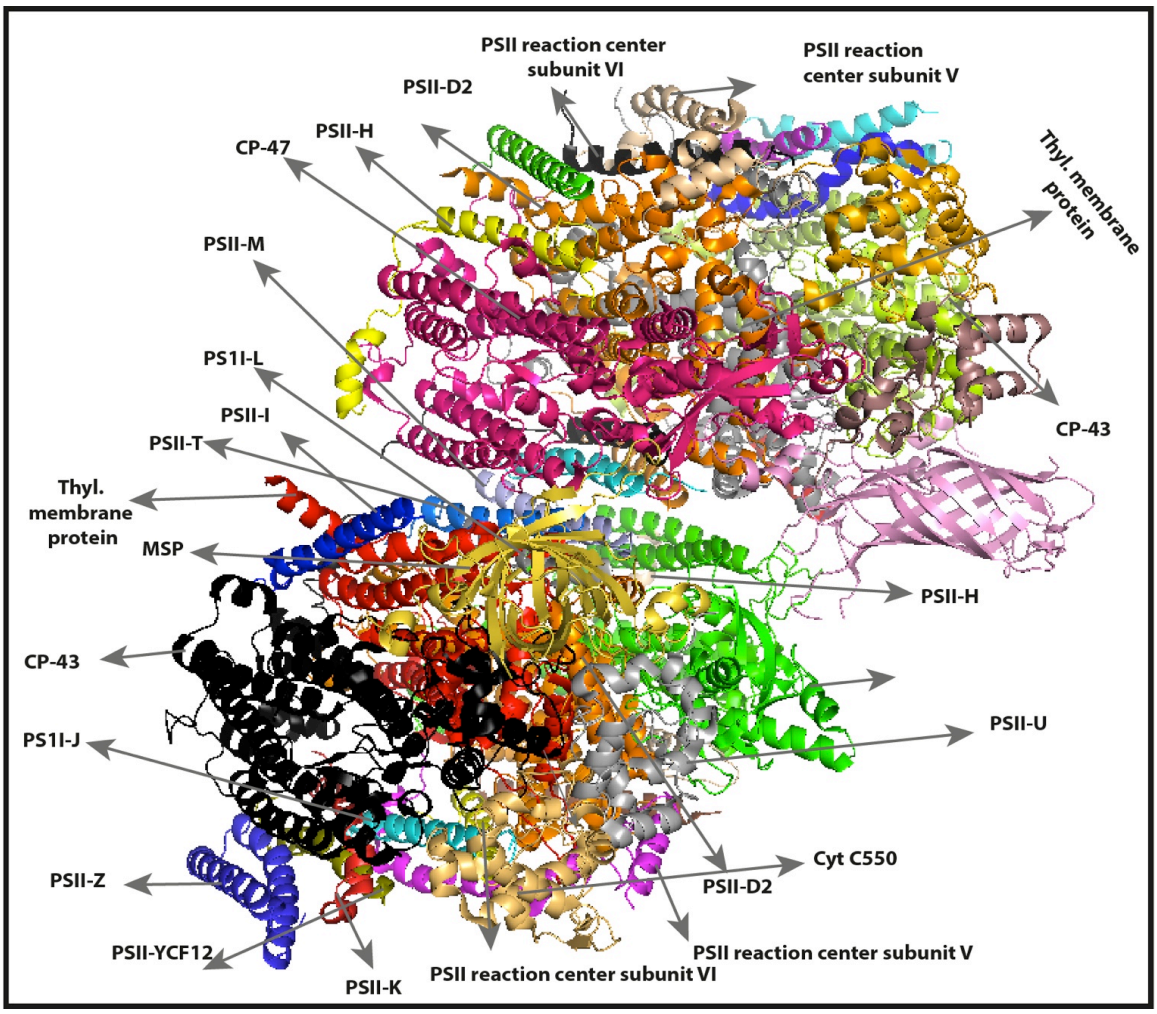


Figure 1.6 Overall structure of PSII dimer at 1.9 Å resolution.

A detailed illustration of subunits of crystallographic model of PSII dimer isolated from T.vulcanus at 1.9 Å resolution. The PDB accession code for this structure is 3WU2. The ribbon structure has been generated using pyMol.

The photosynthetic reaction centre of PSII is the redox active component of the PSII complex and consists of the D1 and D2 proteins, cyt b_{559} and PsbI. It is flanked by the inner antenna proteins

(CP43 and CP47) and on its luminal surface binds the oxygen evolving complex containing the water oxidizing Mn cluster and the 33, 23 and 17kDa proteins in higher plants.

Collectively these and a set of low molecular weight membrane proteins make up the PSII core, which *in vivo* typically functions as a dimer (Boekema, Hankamer et al. 1995, Hankamer, Morris et al. 1999, Hankamer, Morris et al. 2001). The PSII core dimer is associated with an outer antenna, which in higher plants consists to the minor antenna proteins (CP24, CP26 and CP29) as well as the major LHCs 1, 2 and 3. The outer antenna dynamically adjusts to changing light levels and the redox state of the cell. The PSII supercomplex which consists of the PSII core dimer, CP26, CP29 and two LHCII trimers is commonly found in both higher plants and microalgae but a whole host of larger 'mega complexes' containing CP24 and additional LHC trimers have also been reported (Dekker and Boekema 2005). The PSII core dimer (Figure 1.6) isolated from Cyanobateria has been resolved at resolutions of 3.5-1.9 Å (Ferreira, Iverson et al. 2004, Guskov, Kern et al. 2009, Umena, Kawakami et al. 2011). X-ray crystallography over many years has gradually revealed the atomic structure of the cyanobacterial PSII providing a detailed insight into the structural basis for the mechanism of oxygen evolution (Suga, Akita et al. 2015).

1.3.3.2 Cytochrome b_6f

The Cytochrome b_6f complex is located in the thylakoid membrane in chloroplasts. It is composed of two sets of eight protein subunits (Whitelegge, Zhang et al. 2002, Baniulis, Yamashita et al. 2009). A similar X-ray structure (Figure 1.7) was resolved from green algae *(C. reinhardtii)* at 3.1 Å resolution (Stroebel, Choquet et al. 2003). It shuttles electrons between PSII and PSI by transferring them between two electron carriers (plastoquinone and plastocyanin) while releasing two protons from the stroma into the lumen. Biologically, during this electron transport process, protons are transferred from the stroma to the lumen, which contributes to the H^+ gradient that drives ATP synthesis. $Cytb_6f$ is also involved in cyclic electron flow around PSI, which is important to balance the rate of ATP and NADPH production to cellular requirements (Munekage, Hashimoto et al. 2004). Crystal structures have been resolved and deposited for the native b_6f complex (Yamashita, Zhang et al. 2007), and the complex in the presence of the quinone analogue inhibitors DBMIB (Yan, Kurisu et al. 2006), NQNO and tridecyl-stigmatelin (Yamashita, Zhang et al. 2007) with pdb accessions: 2E74, 2D2C, 2E75 and 2E76.

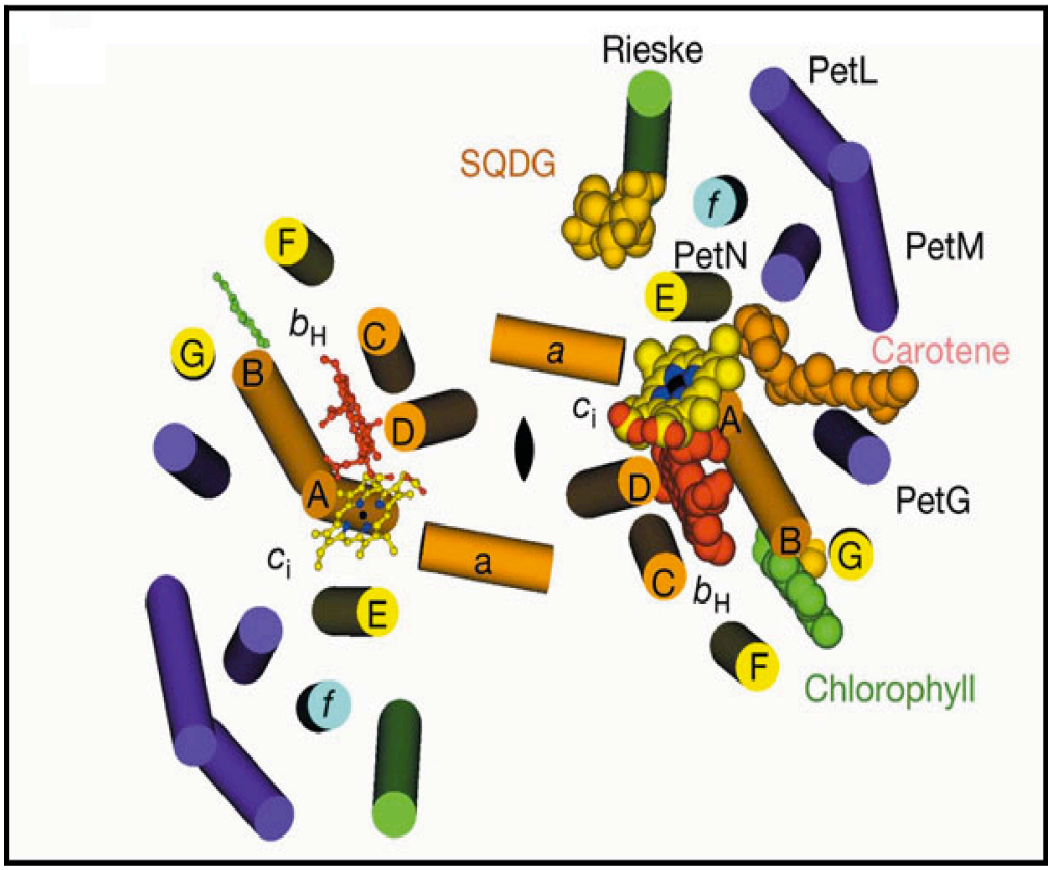


Figure 1.7 X-ray structure of cytochrome b_6f at 3.1 Å resolution.

Stromal view of X-ray structure of cytochrome b_6f isolated from *C. reinhardtii. Cytochrome b_6 represented in* orange helices, small subunits are in purple, subunit IV (E, F and G) in yellow helices represented by cylinders (Stroebel, Choquet et al. 2003).

1.3.3.3 Photosystem I

Photosystem I (PSI) is a multisubunit membrane protein complex consisting of the PSI core and its associated light-harvesting complex (LHCs). The PSI core is responsible for light driven charge separation and electron transfer from plastocyanine to ferrodoxin. The redox active reaction centre consists of PS1-A and PS1-B is evolutionarily related to the PSII core components D1, D2, CP47 and CP43 and is tightly associated with a number of additional proteins, which collectively make up the PSI core, which couples to the peripheral antenna system (Barber, Nield et al. 1997).

The crystal structure of plant PSI has been resolved to 3.4 Å, revealing 17 protein subunits (Amunts, Drory et al. 2007) and 173 chlorophylls, out of which the largest subunits, PsaA and PsaB, form a symmetry-related pseudo-dimer near the center of the complex. Recently an improved crystallographic model was obtained for plant PSI yielding and improved electron density map and leading to the identification of a previously unresolved subunit PsaK (Amunts, Toporik et al. 2010). This new crystal structure reveals the location of and interactions among 17 protein subunit and 193 non-covalently bound photochemical cofactors. Figure 1.8, shows a further improved plant PSI crystal structure that confirms the positions of LHCs and subunits as described by Amunts and

colleagues (Mazor, Borovikova et al. 2015). In another study, the PSI complex of *C. reinhardtii* reveals the main properties of PSI-LHCI super complex. A projection map at 15 Å resolution was obtained by electron microscopy (Drop, Webber-Birungi et al. 2011). This new structure represents one of the largest purified PSI-LCI supercomplexes containing both the core complex and nine Lhca antenna proteins.

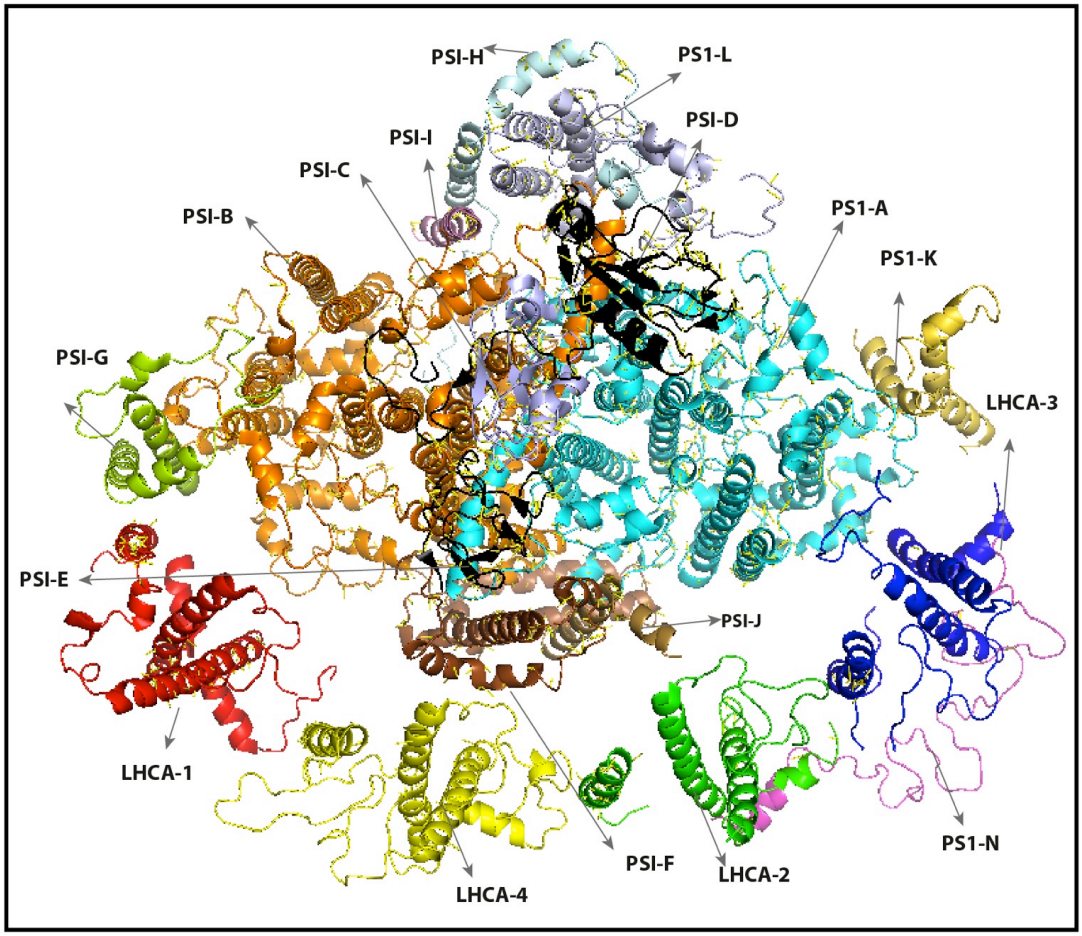


Figure 1.8 Crystal structure of plant PSI at a resolution of 3.0 Å resolution.

Crystal structure of photosystem-I from plant at a resolution of 3.0 Å. Each subunit has been color-coded. The PDB accession code for this structure is 4RKU. The ribbon structure has been generated using pyMol.

1.3.3.4 ATP synthase

The proton gradient across the thylakoid membrane is used to drive ATP synthesis and in so doing combines adenosine diphosphate (ADP) and inorganic phosphate (P_i) to generate ATP. ATP synthases from plants and photosynthetic microalgae (F_1F_0-ATP synthase) belong to the family of F-type ATP synthase enzymes, which are located in between the thylakoid membrane and inner mitochondrial membrane. They are composed of 3 main structural domains. The hydrophobic F_0 domain responsible for proton translocation, the extrinsic F_1 domain that catalyses ATP production and a stalk region connecting them (Oster and Wang 2000, Dekker and Boekema 2005). The mechanical energy generated during the H^+ driven rotation of F_0, is transferred via the central stalk to F_1 to drive ATP synthesis. ATP was discovered by a German chemist Karl Lohmann in 1929

(Saraste 1999). A few years later ATP was synthesized chemically and in 1948, research showed that ATP could be considered the universal energy carriers in cells. In 1997, Paul D. Boyer and John E. Walker elucidated the mechanism of the formation of adenosine triphosphate (ATP) and shared the Noble Prize (Saraste 1999). Boyer and his colleagues proposed how ATP was formed from ADP and phosphate, while Walker and his colleagues established the structure of ATPase and verified its mechanism (Boyer and Walker 1997). The 3D structural information is now available for two respiratory enzyme complexes (F_1F_0) and the catalytic component of ATP synthase at atomic resolution. The structure of the F_1- ATPase (ATP synthase from mitochondria) has been determined at 2.8 Å resolution by X-ray crystallography for the bovine mitochondrial enzyme (Abrahams, Leslie et al. 1994). The structure of the entire ATP synthase from bovine heart mitochondria was determined by single particle electron cryo-microscopy (Rubinstein, Walker et al. 2003) at a resolution of 32 Å, from which the location of all three domains could be inferred (F_0, $F_1$and a peripheral stalk). The low resolution is due to the c1 symmetry of the complex and its structural flexibility. Even after extensive research, a complete high resolution structure of the F_1F_0-type ATPase complex structure (Gibbons, Montgomery et al. 2000, Groth and Pohl 2001, Weber 2007) has not been obtained yet, but all three domains (F1-ATPase at 3.2 Å (Groth and Pohl 2001), central stalk at 2.4 Å (Gibbons, Montgomery et al. 2000)) are available in discrete parts and higher resolution structure of other classes of ATP synthase (Figure 1.9), which bear some resemblance to the F-type ATP synthase, are available such as reported by (Zhao, Benlekbir et al. 2015).

Despite the significant advances described above and a global effort to elucidate the atomic resolution structure and function of the photosynthetic machinery, our atomic level understanding is still limited. Questions include: What is the 3D organisation of the photosynthetic machinery in the native thylakoid membranes at a given point in time? How does the photosynthetic machinery change dynamically in response to changing environmental conditions and what of the key structure-function relationships at the atomic level? What is the purpose of having 20 different LHC proteins associated with photosystems II and I and how do their roles differ? What are the key structure-function relationships between lipids and proteins in the 3-D organization of the thylakoid membranes? What can we learn from natural photosynthetic systems in terms of the design of future solar fuel technologies? These and many other questions remain under active investigation.

Methods for elucidating high-resolution 3D structures of cells, sub-cellular components, membranes and proteins constitute great assets of modern biology because they help us to understand the structure-function relationships of unique biological entities, from macromolecular complexes to subcellular organelles and entire cells. Structural analysis of biological components can largely be achieved by employing different variants of electron microscopy.

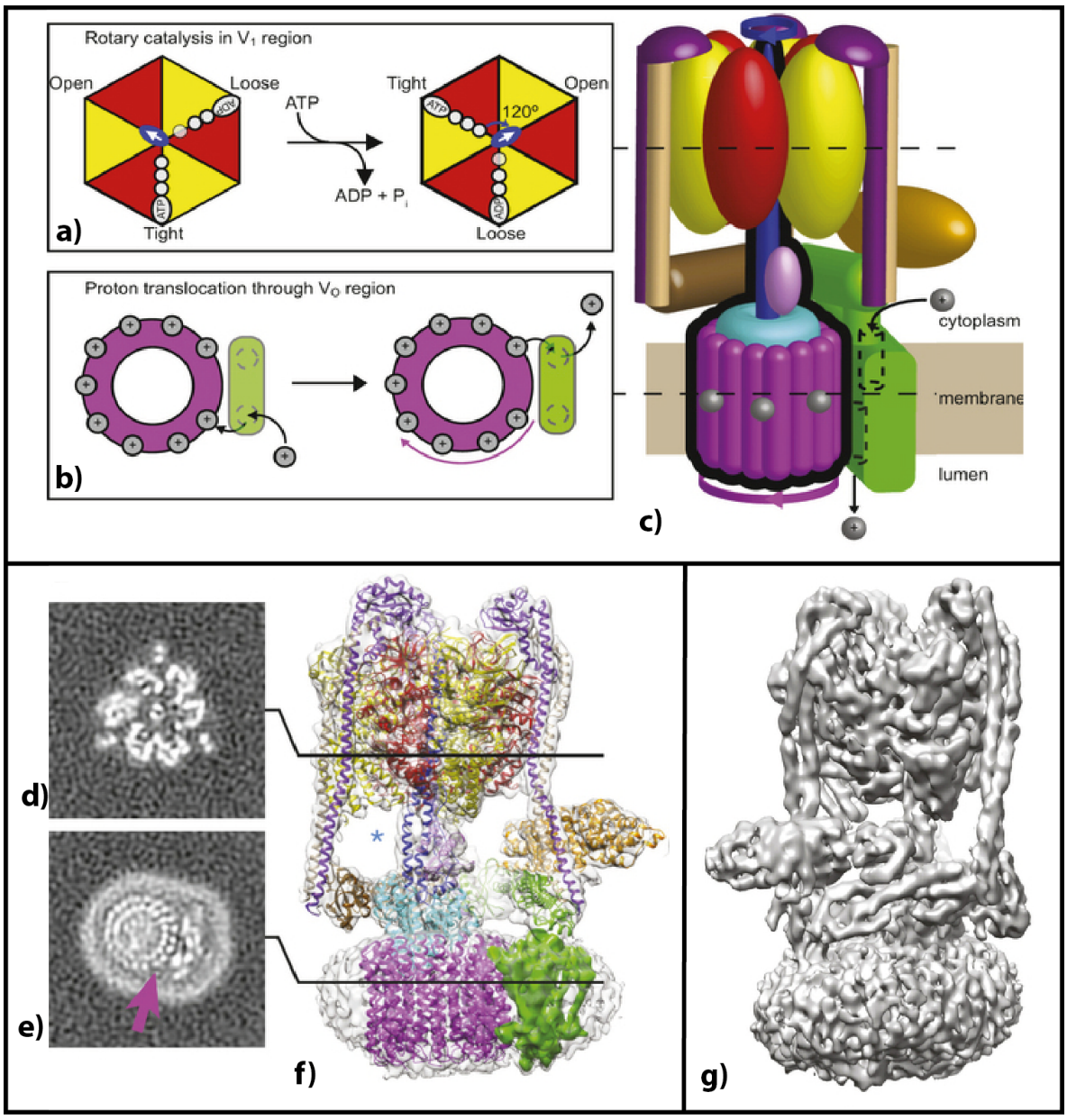


Figure 1.9 The structure of V-ATPase synthase from *S. cerevisiae.*

a) Rotary catalysis in V1 region. b) Proton translocation through Vo region. c) The structure of ATP synthase, which manufactures ATP using ADP and Phosphate (p). d-e) Cross-sections representing views to the rotational state of V1 (upper) and Vo (lower) regions. f) Surface rendering view of 3D map and docking of atomic structures. G) 3D molecular map of ATP synthase determined by single particle analysis . Image reproduced from (Zhao, Benlekbir et al. 2015)

1.3.4 The vision: 3D atomic model using multiscale datasets

Despite significant advances in photosynthesis research, resolving the architecture of the photosynthetic machinery in its 3D membrane architecture remains an ambitious challenge.

More broadly, one of the greatest challenges of cellular and structural biology is to develop approaches to resolve whole cells to increasing resolution and ideally to the atomic level, as this would deliver unprecedented insights into the complex and dynamic interplay between organelles, subcellular structures, macromolecular assemblies and membrane proteins in atomic detail (Figure 1.10). Currently no single structural technique can achieve this aim, but the use of a set of

biophysical techniques spanning the atomic to micron range (e.g. optical microscopy, cellular and single particle electron tomography, biomolecular NMR, X-ray and electron crystallography and molecular docking) offers the potential to yield 'pseudo-atomic' resolution 3D cellular atlases based on nested, multiscale datasets (Alber, Forster et al. 2008).

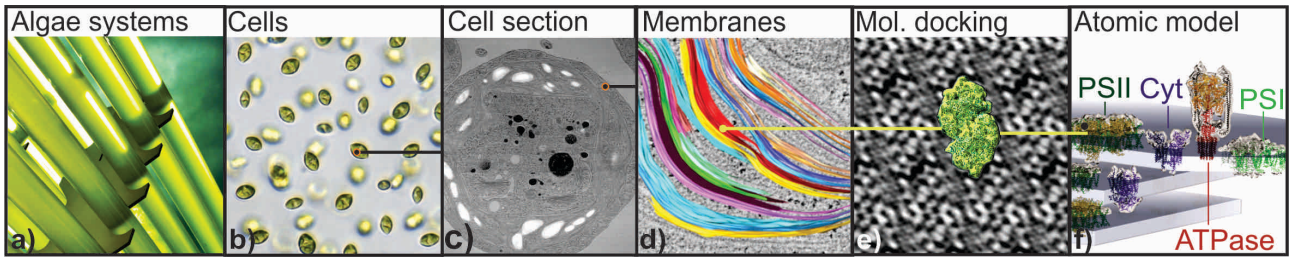


Figure 1.10 3D pseudo atomic 3D atlas of photosynthetic machinery

Representation of multiscale data to achieve pseudo-atomic 3D atlas of photosynthetic machinery. (a) Whole cells grown in laboratory scale photo-bioreactor. Individual cells (b) were being imaged (c). Thylakoid membranes are then magnified, and segmented (d) to locate the photosynthetic super complexes. Edge detection and segmentation tools delineate the 3D contours of individual complexes to resolve individual photosystems (e) at atomic level to construct a molecular resolution 3D atlas (f) of photosynthetic machinery. (Image courtesy of Dr. Ben Hankamer)

The techniques of X-ray and electron crystallography, NMR, single particle analysis and molecular docking are well established in terms of their ability to recover atomic resolution (10^{-10} m) detail for soluble proteins, membrane proteins and macromolecular assemblies (Amunts, Drory et al. 2007, Alber, Forster et al. 2008, Hellmich and Gaudet 2014, Shi 2014). For example, Figure 1.11 shows a set of modelled photosynthetic complexes (Stroebel, Choquet et al. 2003, Standfuss, Terwisscha van Scheltinga et al. 2005, Amunts, Drory et al. 2007, Tokutsu, Iwai et al. 2009, Daum, Nicastro et al. 2010, Umena, Kawakami et al. 2011, Drop, Webber-Birungi et al. 2014), which combine knowledge gained from single particle analysis, electron and x-ray crystallography.

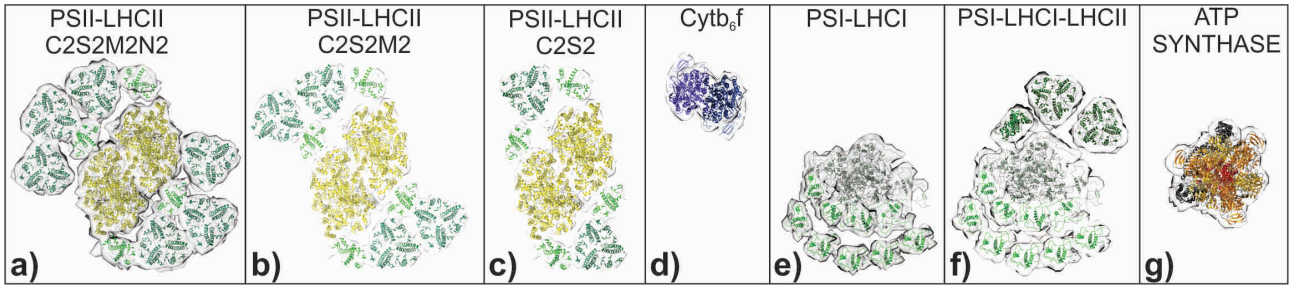


Figure 1.11 High-resolution crystal structures of thylakoid membrane proteins.

(Image courtesy of Dr. Ben Hankamer).

Similarly, advanced optical, super resolution microscopy technologies have enabled the recovery of data in the 10^{-3}m-10^{-9} range (Schermelleh, Heintzmann et al. 2010). However, the routine recovery

of cellular data in the 2-5 nm resolution range, required to resolve macromolecular assemblies and large membrane proteins in their cellular context, can currently only be achieved using tomographic techniques including advanced serial sections scanning electron microscopy and transmission electron tomography. These techniques have in recent years undergone significant improvements in sample preparation and instrumentation (Denk and Horstmann 2004, Liao, Cao et al. 2013, Villa, Schaffer et al. 2013, Amunts, Brown et al. 2014, Kuhlbrandt 2014, Pantelic, Fu et al. 2014, Ramachandra, Bouwer et al. 2014), allowing structural biologists to resolve not only organelle and membrane structures, but cytosolic macromolecular assemblies, cellular motors and the extrinsic densities of large membrane proteins embedded in native membranes in unprecedented detail (Daum, Nicastro et al. 2010, Landsberg, Jones et al. 2011, Liao, Cao et al. 2013, Amunts, Brown et al. 2014, Kuhlbrandt 2014, Pantelic, Fu et al. 2014). The advent of direct electron detectors (DED), single electron counting and super resolution capabilities for electron microscopy has significantly increased obtainable Signal to Noise Ratio (SNR) across the spectral range which are reflected in higher detective quantum efficiency (DQE) (McMullan, Chen et al. 2009). Despite these improvements the accurate and automated detection of contours that define discrete structures of cellular and molecular features remains a computational challenge, as does the high-throughput selective extraction of molecular tomograms for sub-volume averaging and molecular annotation and docking.

In the following section advances in electron microscopy techniques are summarized before proceeding to the core focus of this thesis; the de-noising and segmentation of electron tomograms to facilitate the development of pseudo-atomic resolution models of the future such as those of the photosynthetic machinery, to enable targeted engineering of algae cell lines and the structure guided design of next generation bio-inspired solar fuel systems.

1.4 Electron Microscopy

Electron microscopy (EM) works on a similar principle to that of optical microscopes but uses highly energetic electrons instead of light to obtain the projection image of the sample under observation. Conventional light microscopy has historically been limited to ~ 200nm resolution due to the wavelength of visible light. However, the wavelength of the electron beam is much shorter than of conventional light. EM is thus capable of providing finer details of much smaller objects, with resolution tending to be limited by sample preparation rather than technical limitations of the microscope (Zanetti, Riches et al. 2009, Penczek 2010). In transmission electron microscopy, which is the main focus of this thesis, image contrast is formed by the interaction of the electrons with the sample. The interaction of high-energy electrons with electron dense atoms in the sample can result in energy transfer and the deflection of the incident electrons. This is referred to as elastic scattering

and contributes to signal and image contrast. Electrons can undergo single or multiple scattering events. High-energy electrons that pass through the sample without being deflected are referred to as inelastically scattered electrons and typically increase the background noise. In scanning electron microscopy, when the high energy electron beam interacts with specimen atoms, it typically results in back-scattered (reflected) electrons and secondary electrons (emitted by the surface on excitation) which contribute to signal and image contrast. High-resolution surface information can be obtained from the surface using this approach, but resolution drops off with sample depth.

1.4.1 Types of electron microscope

Several types of electron microscope have been developed for structural biology and these include transmission electron microscope (TEM) and scanning electron microscope (SEM).

1.4.1.1 Scanning Electron Microscope

Scanning electron microscope (SEM) is mostly used as a surface visualisation tool. To create an electron beam a v-shaped cathode is heated up by electric current (Figure 1.12A). A central disk-like anode is installed under the cathode. The primary electrons are accelerated downwards through the annular anode and a series of electromagnetic lenses towards the sample. To image the sample the electron beam is raster-scanned pixel by pixel over the sample and interacts strongly with electron dense atoms (e.g. metals). These atoms absorb the energy and produce reflected/backscattered electrons or secondary electrons. The secondary electron detector captures these secondary electrons and retrieves the structural information from them. The resulting signal is processed to form a grey-scale image on a monitor (Figure 1.12A). Secondary electrons are emitted from locations close to the surface of the specimen. Consequently, SEM can produce very high-resolution images of a sample surface. Reflected/ backscattered electrons come from the surface and from deeper locations within the specimen and consequently produce low resolution images as compare to secondary electron images.

Serial block-face scanning electron microscope (SBF-SEM) enables the collection of automated high-resolution 3D images from fixed biological samples in a reproducible manner. For example, Dr. Winfried Denk who worked on the development of SBF-SEM at Max-Plank-Institute in Heidelberg, analysed the connectivity of axons in the brain in 3D (Denk and Horstmann 2004). Gatan Inc has since commercialised this technology as 'The 3View system'. It consists of a scanning electron microscope coupled with an internal microtome to enable serial section scanning electron microscopy. To achieve this after an image is taken, the microtome slices away a thin section of the surface, to reveal the sample beneath, which is then in turn imaged with the scanning beam. This process is then repeated and can continue in a semi-automated fashion allowing collection of large volumes of 3D data, for computational analysis. The z resolution is largely

controlled by the slice thickness. Whilst this method results in lower resolution reconstruction (~20nm) compared to TEM tomography (~1-4nm), the high throughput semi-automated process of image collection and tomogram reconstruction results in a greatly reduced timeframe for the collection of whole cell 3D data.

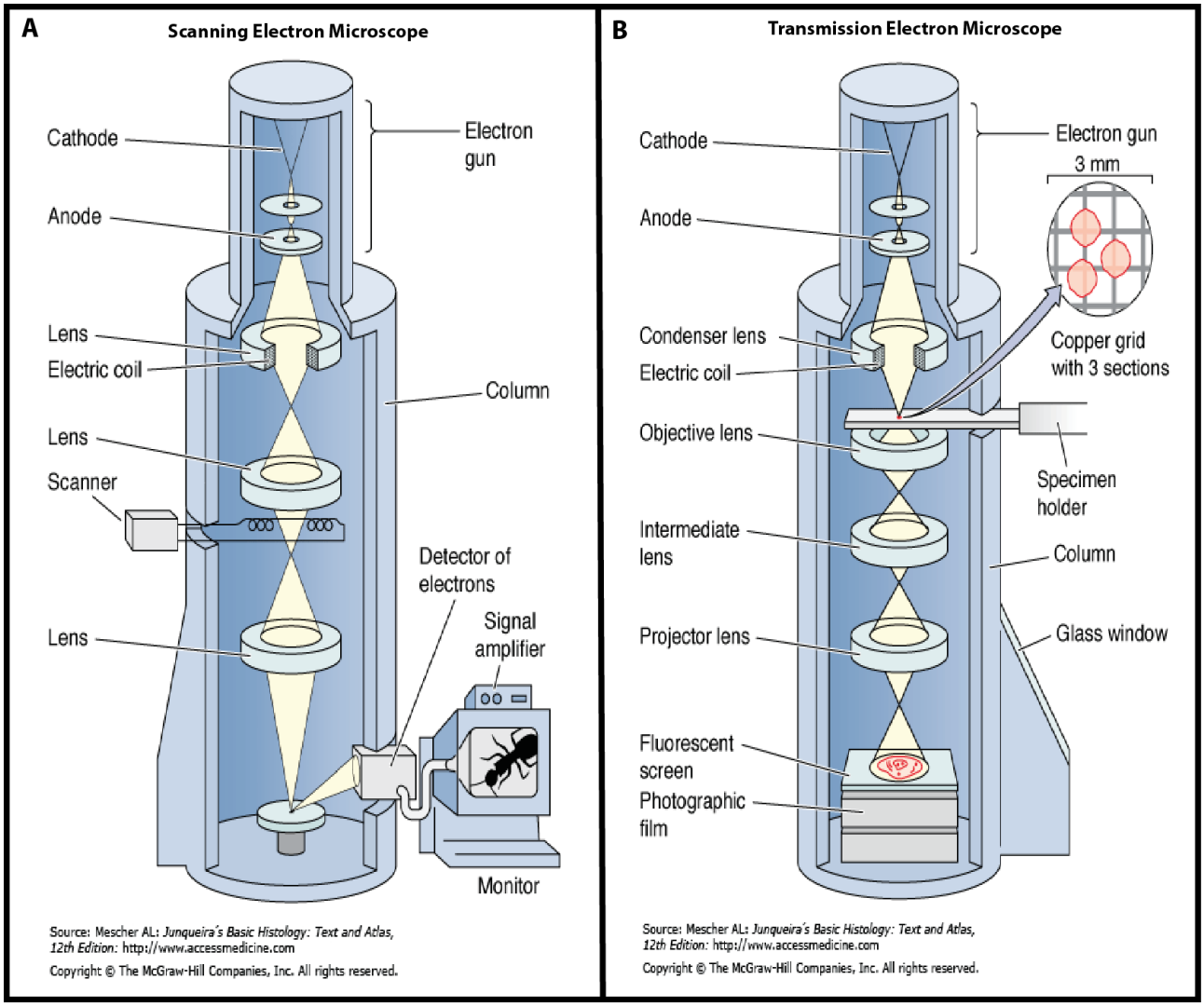


Figure 1.12 Fundamental schematic view of electron microscopes.

The schematic view of Scanning electron microscope (SEM) and Transmission electron microscope (TEM) showing the interaction of electron and samples (Mescher 2009).

1.4.1.2 Transmission Electron Microscope

Transmission electron microscopes (TEM) also uses a high voltage electron beam emitted by a cathode (a very thin metallic filament). The released primary electrons are then subjected to a voltage difference between the cathode and the anode (a metallic plate with a hole in its centre). This electron gun (a source of high energy primary electrons) thereby produces a coherent electron beam, which passes down the electron microscope column in a vacuum through a series of lenses (Condenser lens, Objective lens, Intermediate lens and Projector lens) and the sample to the back focal plane where the image is recorded. The first lens is the condenser lens, which focuses the

electron beam onto the thin specimen. In the case of plastic embedded blocks, samples are typically cut with glass or diamond knives, and the sections transferred to copper support grids for microscopic analysis. This process yields a focused magnified image, based on how many electrons are absorbed by different regions of the specimen, that is then projected through other magnifying lenses onto a fluorescent screen, a CCD camera or direct electron detectors for data collection.

To be imaged in an electron microscope, biological specimens require extensive preparation typically using heavy metal "negative" staining or cryo-preservation. Cryo-preservation techniques require the use of liquid ethane to freeze a specimen rapidly and suspend them in vitreous ice. Cryo-preservation more closely reflects the native state of the specimen due to a lack of heavy metal staining, dehydration artefacts, and enhanced penetration of water (versus stain) into the internal structures. Negative staining employs heavy atoms such as lead and uranium at room temperature and while technically less demanding and providing more contrast, is accompanied by an increase in the inherent noise and a reduction in resolution.

1.4.1.2.1 Sample limitations

To image biological specimens in TEM there are additional crucial factors including purification and preparation techniques that limit the ability to recovery of high-resolution information. Furthermore, biological samples are very sensitive to beam/radiation damage. Therefore, low dose imaging conditions result in extremely noisy tomograms. Low contrast and high noise levels hamper the processing of volumetric data and create difficulties in visualisation and interpretation. Biological samples are damaged by strong exposure to high-energy electrons of the electron beam. In order to avoid structural damage, the electron dose should typically be kept sufficiently low to minimise damage while still yielding sufficient contrast to enable structure determination.

The stability of proteins is another limiting factor in the optimisation of protein preparations. The screening of optimal pH and finding appropriate buffering conditions are important issues that need to be dealt with carefully. Sparse-matrix screening of their thermal unfolding behaviour in the presence of various buffers and small molecules can optimise the stabilising conditions (Chari, Haselbach et al. 2015). Based on such analysis, glycerol gradient/crosslinking approaches (e.g. Grafix, a density gradient centrifugation procedure (Kastner, Fischer et al. 2008)) can be used for size fractionation and fixation of macromolecules by mild chemical fixation. Proteoplex and GraFix plays important role in optimising the stability, homogeneity and solubility of macromolecular complexes (Stark and Chari 2016) for single particle analysis.

Compared to amorphous carbon films, graphene is increasingly used as a TEM support film. It is atomically thin, and consists of light atoms with a highly ordered structure. It is electrically and thermally conductive. Remarkable properties such as markedly reduced phase or amplitude contrast

and high electrical carrier mobility (Pantelic, Fu et al. 2014) make graphene a promising support film for EM.

In recent years such significant improvements in sample preparation and instrumentation have been made that these now allow structural biologists to resolve not only organelle and membrane structures, but cytosolic macromolecular assemblies, cellular motors and the extrinsic densities of large membrane proteins embedded in native membranes in unprecedented detail.

1.4.2 Limiting factors of electron microscope to achieve high resolution

1.4.2.1 Point Spread Function (PSF)

No imaging system is perfect and so will yield an image of the object with some degree of distortion. The PSF is a measure of the impulse response of an imaging system to a point source. The PSF can thus be taken as an indicator of the quality of an imaging system, indicative of the degree of spreading/ blurring from any given point on the object being imaged. Figure 1.13b shows that the distribution of image signal is achieved by convolving the PSF with the object. The limitations of the sensor or its environment, noise filtering, and correction of geometric distortions due to sensors, degrades the image and gives a blurred output. In this example $f(x,y)$ and $h(x,y)$ represent the point object and PSF respectively (Figure 1.13a). The output image $g(x,y)$ can be written mathematically and is described in equation 1.1.

$$g(x,y) = f(x,y) * h(x,y) \quad (1.1)$$

1.4.2.2 Optical Transfer Function (OTF)

The optical transfer function (OTF) is a complex transfer function in an optical system such as a camera, microscope, human eye, or projector. The OTF is composed of two parts, real and imaginary. The first part of the function is responsible for the modulation transfer and is referred to as the modular transfer function (MTF), whereas the imaginary part describes the phase transfer (PhTF). By definition, the OTF is a measurement of the ratio of image contrast to specimen contrast.

OTF = MTF x $e^{i\theta(f)}$, where $e^{i\theta(f)}$ is an imaginary term used to represent PhTF.

1.4.2.3 Modulation Transfer Function (MTF)

The MTF describes the lens performance (light distribution from the image of a point object (PSF)) produced by the optical system. In simple words, the MTF is the frequency response of the PSF. Mathematically, the frequency response is given by the FT of the PSF as shown in equation 1.2.

$$M(f) = \frac{G(f)}{F(f)} \tag{1.2}$$

Where $M(f)$ represents the MTF, $G(f)$ and $F(f)$ are frequency domain representations of the output and input image respectively. In Figure 1.13, two output images are shown.

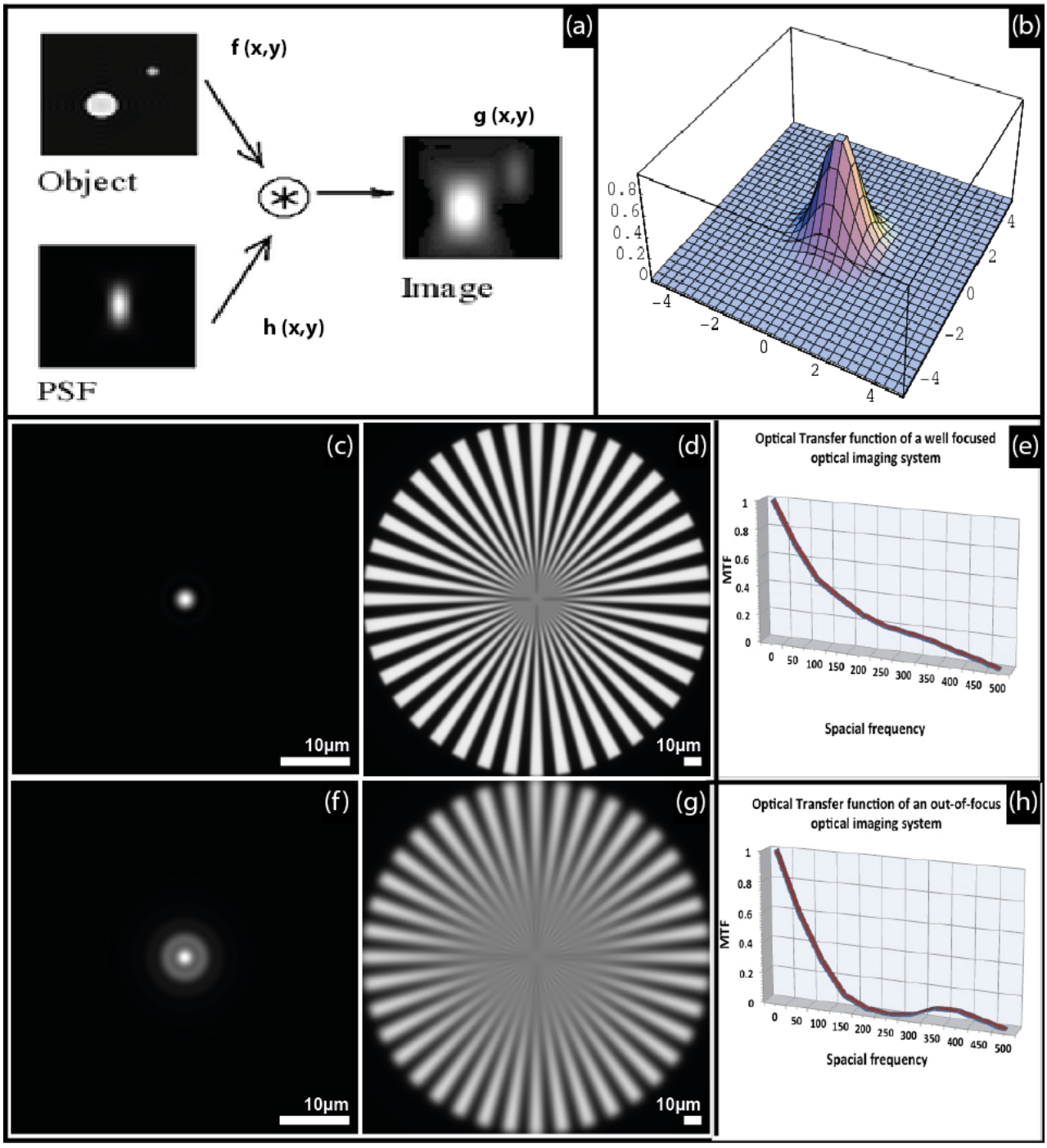


Figure 1.13 Illustration of PSF and MTF and its effect on image.

The convolution of point-spread function h (x,y) with object f (x,y) results in blurring point image g(x,y). b) Image is blurred and degraded due to noise by optical system, and can be represented as Gaussian distribution. c & f) The modulation transfer function is performed on images(point source) generated through an ideal system and imperfect lens system respectively. d & g) images of spoke targets . e) Response of MTF of a well focused, and out of focus optical imaging system.

The top one represents an ideal system while the bottom one shows an imperfect lens system. For the ideal system it can be seen that information is captured across the resolution range from low to high frequency (Figure 1.13; c-e). For a non-ideal system, the output image loses high frequency information and cannot resolve the high-frequency stripes towards the center of the image (Figure 1.13; f-h). The MTF therefore results in a frequency response that gradually decreases from low to high frequency. In the ideal system, the drop-off is relatively slow compared to the non-ideal system. Thus, while the PSF expresses system response in the spatial domain, the MTF describes this in the frequency domain.

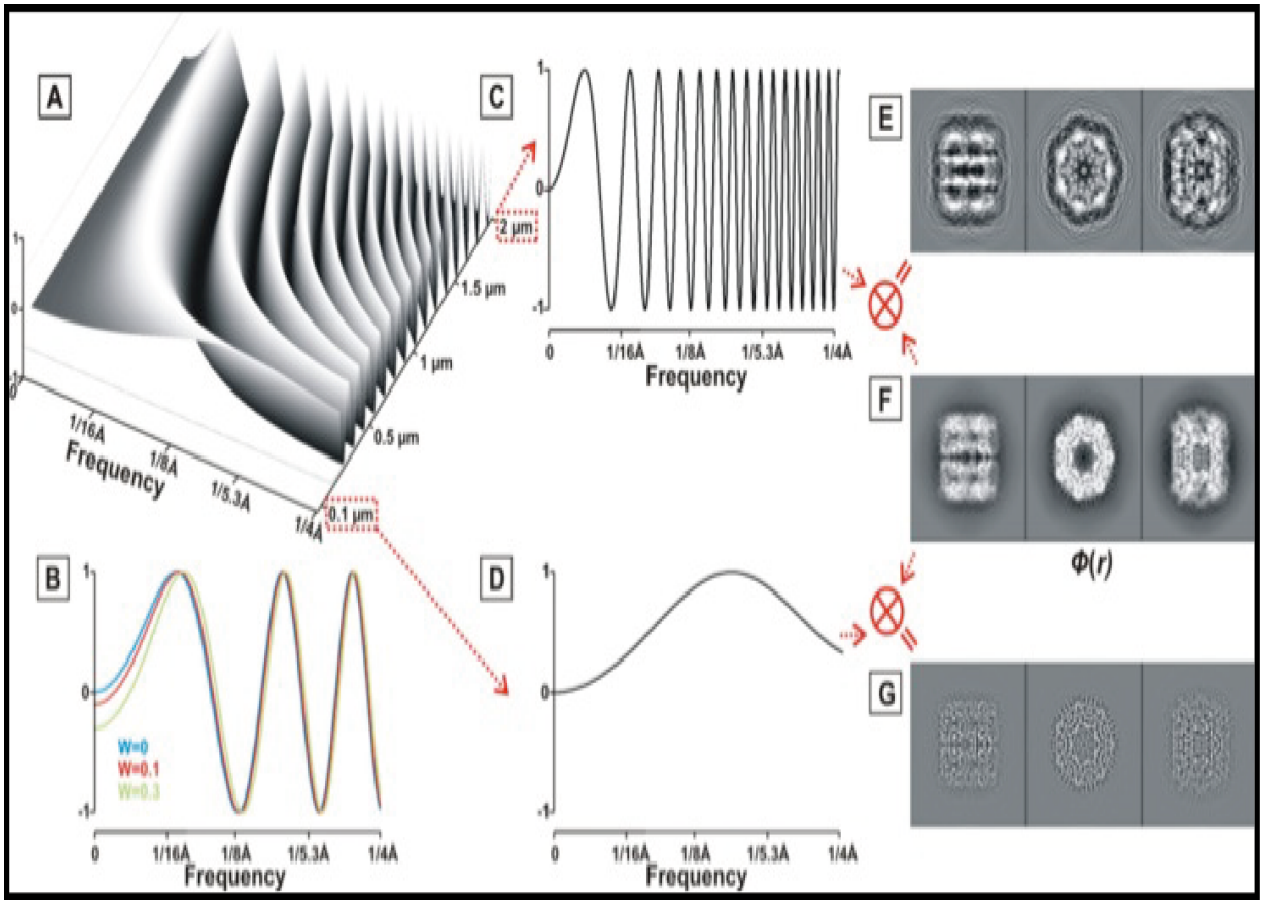


Figure 1.14 The effect of the contrast transfer function on image formation.

A). CTF plotted as a surface representation to show its dependence on defocus. B). Result of increasing the amplitude attenuation term (w), C. The CTF profile at a defocus of 3µm. D). The CTF profile at a defocus setting of 0.1 µm. E). The result of convolving the CTF depicted in C to the pure projection data in F) Pure projection data of the GroEL model, generated from the published 6 Å structure (Ludtke, Chen et al. 2004) G). The result of convolving the CTF depicted in D to the pure projection data in F. (Woolford 2007)

1.4.2.4 Contrast Transfer Function (CTF)

In EM, the Contrast Transfer Function (CTF) describes the relationship between the object and its image in TEM, essentially providing a description of distortions introduced due to imperfect image formation by the microscope. As shown in Figure 1.14, the CTF modulates the resultant image in a spatial frequency-dependent manner. The CTF of an electron microscope is dependent on the level

of defocus and this can be taken advantage, in order to achieve sufficient contrast so that low contrast biological specimens can be visualised. It is also used for evaluation and comparison of performances of different microscopes.

1.4.2.5 Detective Quantum Efficiency (DQE)

DQE is a measure of the combined effects of signal and noise performance (i.e. SNR) of an imaging system across the resolution (frequency) range. McMullan defines DQE as the square of the ratio of the output SNR to the input SNR as shown in equation 1.3 (McMullan, Chen et al. 2009).

$$DQE = \frac{SNR_o^2}{SNR_i^2} \tag{1.3}$$

Here SNR_o^2 and SNR_i^2 represents the output and input SNR respectively. Compared with film or screen imaging, a digital detector with high DQE has potential to deliver significant object-detectability improvements at an equivalent dose, or to permit object detectability comparable to film at a reduced dose. An example is shown in Figure 1.15a that compares the performance of film, and few direct electron and scintillator-based detectors (US4000, F416, Falcon I, Falcon II, DE-12, the K2 summit (counting) and K2 Summit (super resolution)) (Ruskin, Yu et al. 2013).

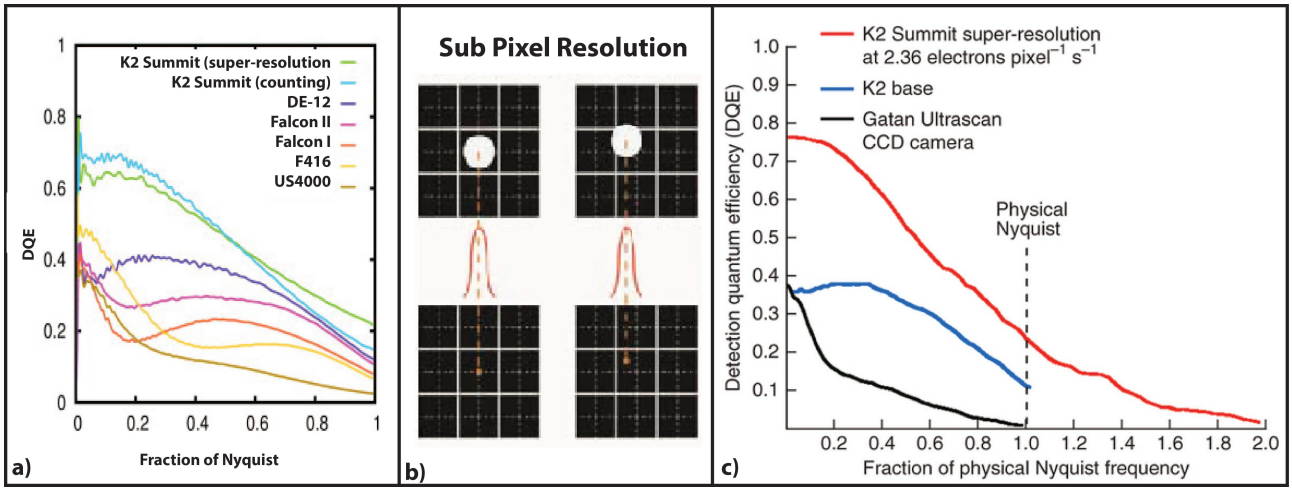


Figure 1.15 DQE of Direct electron detectors and Scintillator based detectors.

a) DQE curves for the DEDs and scintillator-based detectors (for 200keV electrons). Figure adapted from (Ruskin, Yu et al. 2013). b) Sub pixel resolution (Image courtesy of Prof. Ben Hankamer). c) Comparison of DQEs between K2 Summit camera (red), K2 base camera (blue) CCD camera (the Gatan Ultrascan) in black. Figure adapted from (Li, Mooney et al. 2013)

Each electron event on a DED camera can be presented as a Gaussian distribution (Figure 1.15b), which results in improve accuracy up to sub-pixel resolution. As a result sub pixel accuracy of "counting mode' allows a physical 4k x 4k camera to read out an effective 8k x 8k image. This enables the K2 summit to perform at its best at recovering low frequency information (beyond the physical Nyquist limit (Figure 1.15c)), which is critical for the detection of small molecules and

also for particle alignment. Furthermore the K2 Summit also has the highest DQE at high resolution (Li, Mooney et al. 2013).

1.4.3 Advances in electron microscopy instrumentation

Electron microscopy is currently undergoing a 'resolution revolution' (Kuhlbrandt 2014) due to a number of fundamentally important advances at the level of imaging, data capture and image processing. Collectively these are resolving the ultrastructure of cells and their constituent protein complexes ever more quickly and to higher resolution. For example, cryo-focused ion beam (Cryo-FIB) milling is now making it possible to resolve the ultrastructure of whole cells to ~4nm resolution in vitreous ice (Engel, Schaffer et al. 2015). Another key improvement is Cs aberration correctors that correct spherical aberrations of the lens system of transmission electron microscope resulting in improved Contrast Transfer Function (CTF). Phase plates contribute to phase contrast upon formation of a phase contrast image by generating a phase shift of the scattered electrons in microscope. The integration of phase plates (Danev, Buijsse et al. 2014, Asano, Fukuda et al. 2015, Danev and Baumeister 2016) into advanced electron microscopes increases the contrast of cellular densities recoverable by cryo-transmission electron microscopy into the sub-nanometer range, thereby improving the visualisation of macromolecular assemblies and ordered subcellular structures in a near-native environment (Asano, Fukuda et al. 2015). Their resolution in turn is being enhanced further by invoking the idea of averaging of aligned image stacks through the use of sub-volume averaging programs (e.g. Dynamo (Castano-Diez, Kudryashev et al. 2012), PyTom (Hrabe, Chen et al. 2012). Furthermore state-of-the-art direct electron detectors greatly reduce the impacts of sample charging and drift on the captured image, increasingly making atomic resolution structure determination by single particle analysis (Allegretti, Mills et al. 2014) a reality not only for viruses and macromolecular assemblies, but also for membrane proteins smaller than 500kDa (Li, Mooney et al. 2013, Liao, Cao et al. 2013, Lyumkis, Julien et al. 2013, Bartesaghi, Matthies et al. 2014, Campbell, Veesler et al. 2015).

Despite the high theoretical resolution achievable by electron microscopes (a typical 200keV electron has a wavelength of ~2 pm), imaging protein structures at atomic resolution has been difficult to achieve, mainly due to the sensitivity of biological samples to electron beam damage, sample drift and poor image contrast at lower "safe" electron doses. The performance of CCD cameras is adequate when electron energies are between 80-120 keV, but when one goes beyond this range, performance starts to degrade and so does image quality. As an example of this, CCD cameras are inferior to film recording, due to their lower DQE and reduced dynamic range, but have been preferred to film for reasons of convenience. However, improved SNRs throughout the

resolution range, relative to both CCDs and film cameras can now be attained using improved direct electron detection (DED) technology, an advancement which has revolutionised cryo-EM because of its imaging properties (Figure 1.15). The advancement of improved DED cameras has shown potential to overcome these problems with unprecedented success and now has enabled the possibility of determining near-atomic resolution bio-molecular structures (Liao, Cao et al. 2013, Amunts, Brown et al. 2014).

Based on these advances, structural biology is now entering the exciting era of resolving atomic resolution 3D atlases of cells based on the nested docking of multiscale datasets (from whole cell tomograms to atomic resolution protein structures). Tomograms, upon segmentation, provide structural constraints to dock atomic resolution structures of large macromolecular assemblies and protein complexes into cellular tomograms (Davies, Strauss et al. 2011). Accurate, high-throughput cell segmentation capability is therefore a critical link between advanced electron tomography and both high-throughput sub-volume averaging and precise molecular docking.

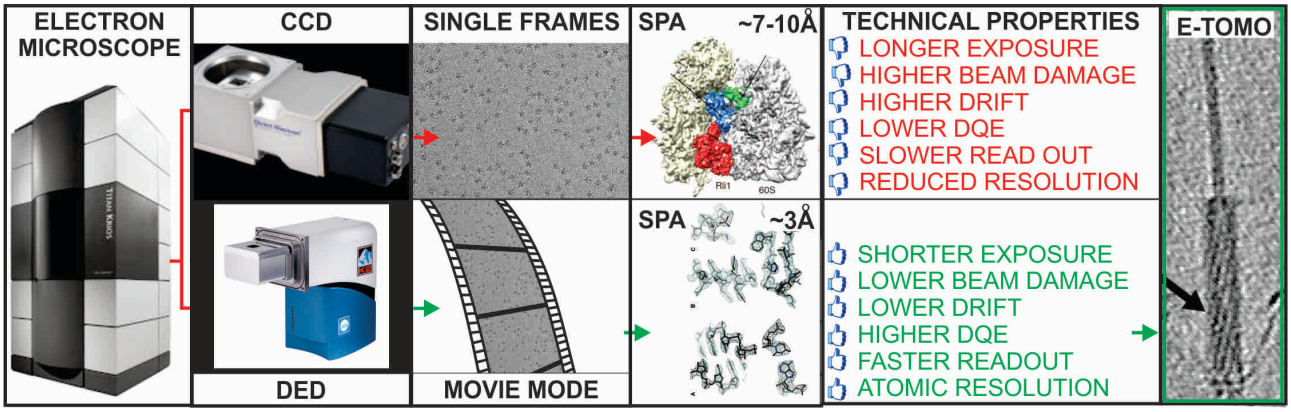


Figure 1.16 Comparison of CCD and DED camera systems and their imaging properties.

(Image courtesy of Dr. Ben Hankamer)

Despite the high theoretical resolution achievable by electron microscopes (a typical 200keV electron has a wavelength of ~2 pm), imaging protein structures at atomic resolution has been difficult to achieve, due mainly to sensitivity of biological samples to electron beam damage, sample drift and poor image contrast at lower "safe" electron doses. The performance of CCD cameras is adequate when electron energies are between 80-120 keV, but when one goes beyond this range, performance starts to degrade and so does image quality. As an example of this, CCD cameras are inferior to film recording, due to their lower DQE and reduced dynamic range, but have been preferred to film for reasons of convenience. However, improved SNRs throughout the resolution range, relative to both CCDs and film cameras can now be attained using improved direct electron detection (DED) technology, an advancement which has revolutionise cryo-EM because of its imaging properties (Figure 1.16).

1.4.3.1 Comparison between CCD and DED

CCD cameras have been extensively used in last decade for acquiring EM data. The main advantage of DEDs is that they have a higher DQE (Figure 1.16c), throughout the resolution range. This helps with recovery of high-resolution information (meaning in theory less particles are needed to get to high resolution) but also improves contrast.

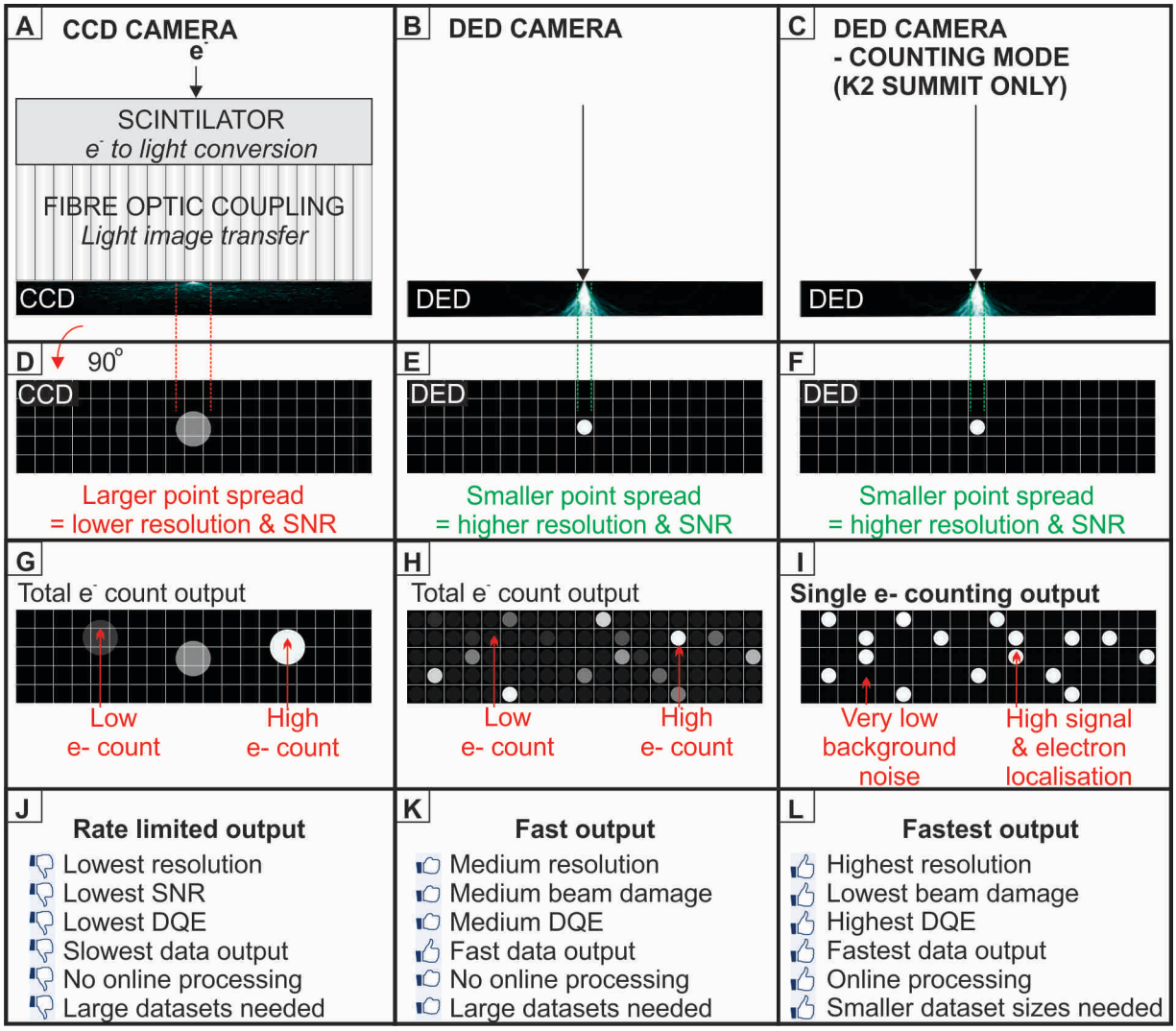


Figure 1.17 Comparison between CCD and DED Camera (counting mode).

Image courtesy of Dr. Ben Hankamer.

The improved DQE is largely due to the reduced PSF, a product of the fact that DEDs are thinner – CCDs capture electrons in an upper scintillator layer, transmit the optical image through a fibre optic coupling and then convert back to an electrical signal, while DEDs directly record incoming electrons on a single, back-thinned sensor (Figure 1.17; A-C). The reduced thickness improves the PSF (Figure 1.17; D-F), while DQE is also improved by the reduced noise as the two photo conversion steps are omitted. With improved contrast, shorter exposures can be used. This reduces sample exposure and thus image degradation as a result of radiation damage.

DEDs also operate at extraordinarily high readout rates (20-400fps). The combination of high readout rates and shorter necessary exposure times means that exposures can be recorded as stacks of images (or "Movies") with the individual frames corrected for things like beam induced motion before summing them together. This further improves image quality.

Finally, the high readout rate and accurate matching of the physical pixel size of a DED detector to the PSF of the incident electrons allows DEDs to be operated in counting mode. Here, the device records each and every incident electron as it hits the DED chip, meaning that incident electron events are normalized and spurious signals can be filtered out. The ability to detect individual electron events also lends itself to super-resolution imaging, where the location of incident electrons is mapped with sub-pixel accuracy (Figure 1.17; G-I) – generating an output image twice the size of the DED array in both X and Y dimensions.

Table 1.1 summarizes the resolution, frame speed, pixel size, MTF and DQE of all three detectors as studied by (McMullan, Faruqi et al. 2014). Less readout noise and high DQE makes K2 summit best candidate to be used in electron tomography.

Table 1.1 Physical properties of the detectors.

Detector	Sensor size	Pixel size	Readout speed	Resolution	MTF	DQE
DE-20	5120×3840	6.4 µm	25 fps	5 Mpix	Medium	Lower
Falcon-II	4096×4096	14.0 µm	18 fps	>14.2 Mpix	Lower	Medium
K2 Summit	3838×3710	5.0 µm	400 fps	14.2 Mpix	Highest	Highest

1.4.3.2 Summary

CCD cameras have low resolution and SNR, inferior DQE and slow readout rates. Also these cameras do not provide online processing capacity. All these factors effect on the speed and quality of structure determination. Comparatively, DED cameras (Gatan K2 Summit in particular) offer highest resolution, least beam damage, superior DQE and fastest data readout with an advantage of online processing. Apparently, DED yields superior structural details to determine a 3D structure at atomic resolution.

1.4.4 Conventional EM methods to study cellular organization

Structural studies incorporating EM have revealed valuable information detailing the dynamic and complex structure of microalgae and their embedded photosynthetic machinery. Mussgnug *et al (Mussgnug, Thomas-Hall et al. 2007)* used EM to compare two strains of *C. reinhardtii*: stm3 and Stm3LR3. They found that stm3LR3 has limited granal stacking where membrane pairs are loosely

stacked, compared to the extensive granal stacks observed for stm3. Such morphological changes of thylakoid organization were not previously described in the literature (Mussgnug, Thomas-Hall et al. 2007) and these findings highlighted the need to determine the high-resolution 3D structures of complex proteins to elicit their individual functions in processes like membrane stacking, for example.

In order to obtain pseudo atomic 3D reconstructions of the embedded photosynthetic machinery *in situ*, there is a need to obtain detailed multiscale structural data by incorporating ET (cellular scale), SPA (molecular scale) and crystallographic data (atomic scale). Such analysis has the potential to derive the structure of the light harvesting processes of *C. reinhardtii.* SPA allows us to resolve the structures of relatively large, non-crystallized proteins and macromolecular complexes at ~0.5 nm resolution (Iwasaki 2006, Cong and Ludtke 2010). ET allows us to reconstruct, visualize and analyze organelles and subcellular structures at ~4 nm resolution (Frangakis and Forster 2004, Volkmann 2010), whereas high resolution crystal structures can provide information down to ~1 Å resolution. The combination of ET, SPA and crystallographic approaches can thus potentially provide a multiscale structural framework to reveal *in situ* membrane protein structures at the sub-nanometer scale (Trepout, Taveau et al. 2010).

1.4.4.1 Single Particle Analysis (SPA)

SPA has emerged as a powerful technique for atomic resolution protein structure determination, particularly of large multi-protein macromolecular assemblies and membrane protein complexes, which are of fundamental biological importance, but often fragile and difficult to isolate, in particular in the quantities required for crystallography.

It is a methodology used to generate 3D volumetric reconstructions of relatively large and mono-disperse membrane proteins and macromolecular assemblies imaged by either negative stain or cryo-EM. Unlike X-ray or electron crystallography, there is no prior need for crystallization of the specimen of interest. This powerful technique can determine 3D structures of soluble macromolecules under conditions close to their native environment (Cong and Ludtke 2010). Particles are either preserved using heavy metal “negative” staining or cryo-preservation techniques. Cryo-preservation techniques require

the use of liquid ethane to freeze a specimen rapidly in vitreous ice. Because the specimen is in aqueous buffer prior to freezing, cryo-presereved specimens are more likely to reflect the native state of the specimen. Reduced dehydration artifacts and enhanced penetration of water (versus stain) into the internal structure of the molecules are additional advantages. Negative staining employs heavy atoms such as lead and uranium at room temperature and while technically less demanding and providing more contrast, is accompanied by reduced resolution. Ultimately, the

achievement of high-resolution 3D structural information of ostensibly identical and individual nano-scale objects not only depends on specimen preparation, but also on high-resolution image acquisition (Cong and Ludtke 2010). A large number of images (10,000 - 1,000,000) can be digitized either from film or imaged directly via a CCD or DDD camera.

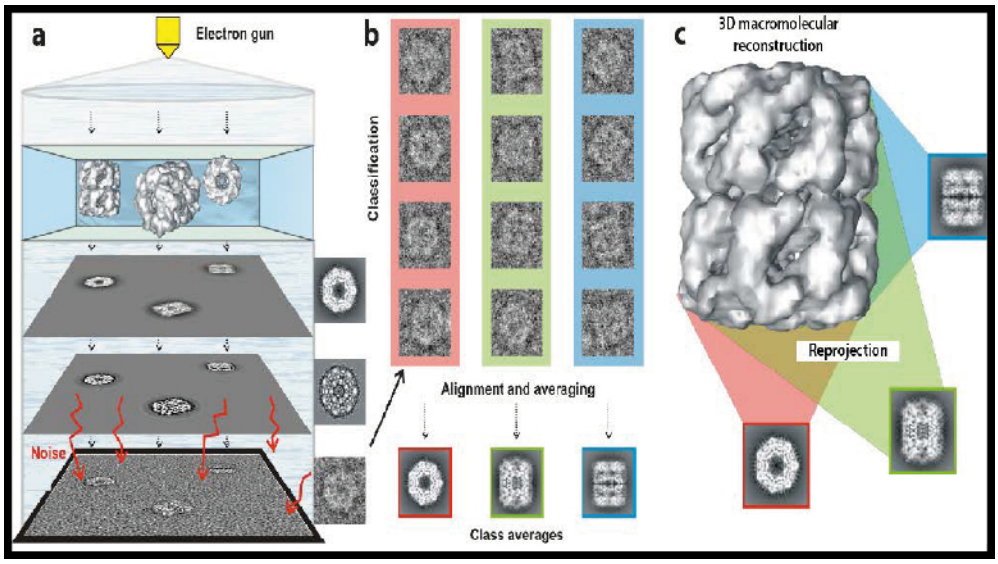


Figure 1.18 Electron Image formation through EM and single particle analysis concept.

a) shows different steps of 2D image formation. Electrons are released from electron gun and the vitrified (purified) sample is exposed through the electron beam. After passage of electron beam thorough sample, it contains all the projection information, which after going through contrast transfer (CTF) and envelope transfer functions produces 2D molecular projections (contaminated with noise) of sample are formed. The final outputs of this procedure are raw digital 2D images. b) 2D projection images are then classified (grouped based on same orientations of sample) together, aligned and averaged. Final class averages represent different views of particle. c) Class averages are then composed together (using back projection technique) to approximate a final 3D structure. Figure adapted from (Woolford, Ericksson et al. 2007).

Many software packages have been developed for single particle reconstruction, including Spider (Shaikh, Gao et al. 2008, Baniulis, Yamashita et al. 2009), IMAGIC (Forster and Hegerl 2007), Xmipp (Sorzano, Marabini et al. 2004), BSoft (Heymann and Belnap 2007, Heymann, Cardone et al. 2008), SPARX (Hohn, Tang et al. 2007) EMAN (Heymann and Belnap 2007), EMAN2 (Tang, Peng et al. 2007), Relion (Scheres 2012), Frealign (Grigorieff 2007). Raw particle images obtained via EM and extracted from larger micrographs are classified based on their apparent orientations. Secondly, members of each class are aligned (transnationally and rotationally) and averaged. Class averages obtained from this step are then used to reconstruct a 3D model, using one of several different approaches that computes the relative orientation of each individual class such that it is most consistent with current 3D model. Iterative refinement of this process continues until a final, convergent 3D reconstruction (Figure 1.18) is obtained.

Furthermore state-of-the-art direct electron detectors greatly reduce the impacts of sample charging and drift on the captured image, increasingly making atomic resolution structure determination by

single particle analysis (Allegretti, Mills et al. 2014) a reality not only for viruses and macromolecular assemblies but also for membrane proteins in the 500kDa range (Li, Mooney et al. 2013, Liao, Cao et al. 2013, Lyumkis, Julien et al. 2013, Bartesaghi, Matthies et al. 2014, Lu, Bai et al. 2014, Campbell, Veesler et al. 2015) and below.

1.4.4.2 Electron Crystallography

Electron crystallography is a structural biology technique most commonly used to determine the high-resolution 3D structures of membrane proteins typically arrayed within to 2D crystals.

In the field of crystallography three main methods are in use: X-ray crystallography, electron diffraction of single crystals and electron crystallography. The advantage of the latter two techniques, is that the interaction of electrons with matter is much stronger than X-rays, such that the loss of information due to damaging, inelastically scattered electrons is up to 1000-fold less than for equivalent X- rays (Henderson 1995). Also, 2D electron crystallography is well suited to studying the structure of membrane proteins embedded in near native lipid bilayers (Kuhlbrandt 2013) but the production of high quality 2D crystals is technically challenging as is the recovery of isotropic 3D data. Two methods are typically used to produce 2D crystals: bilayer and monolayer crystallization.

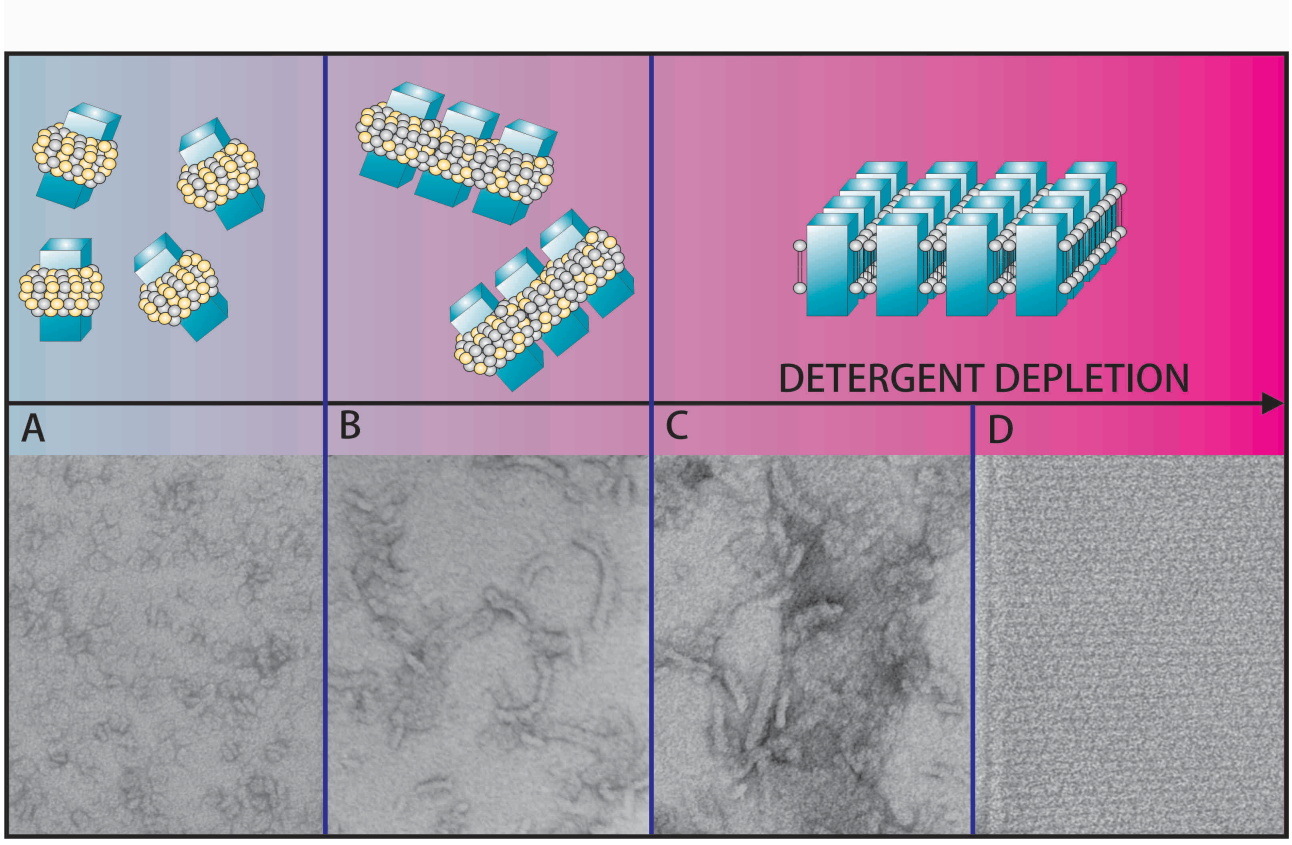


Figure 1.19 Illustration of Bilayer crystallization.

(Image courtesy of Dr. Ben Hankamer)

1.4.4.2.1 Bilayer crystallization

Bilayer crystallization involves the reconstitution of detergent-solubilized membrane proteins (Figure 1.19A; protein blue, detergent yellow) with bilayer-forming lipids (grey). The detergent is then depleted through dialysis or the use of detergent absorbing beads (eg. Biobeads) to induce the formation of protein-lipid aggregates (Figure 1.19B), membrane patches (Figure 1.19C) and in the ideal scenario, eventually 2D crystalline arrays (Figure 1.19D).

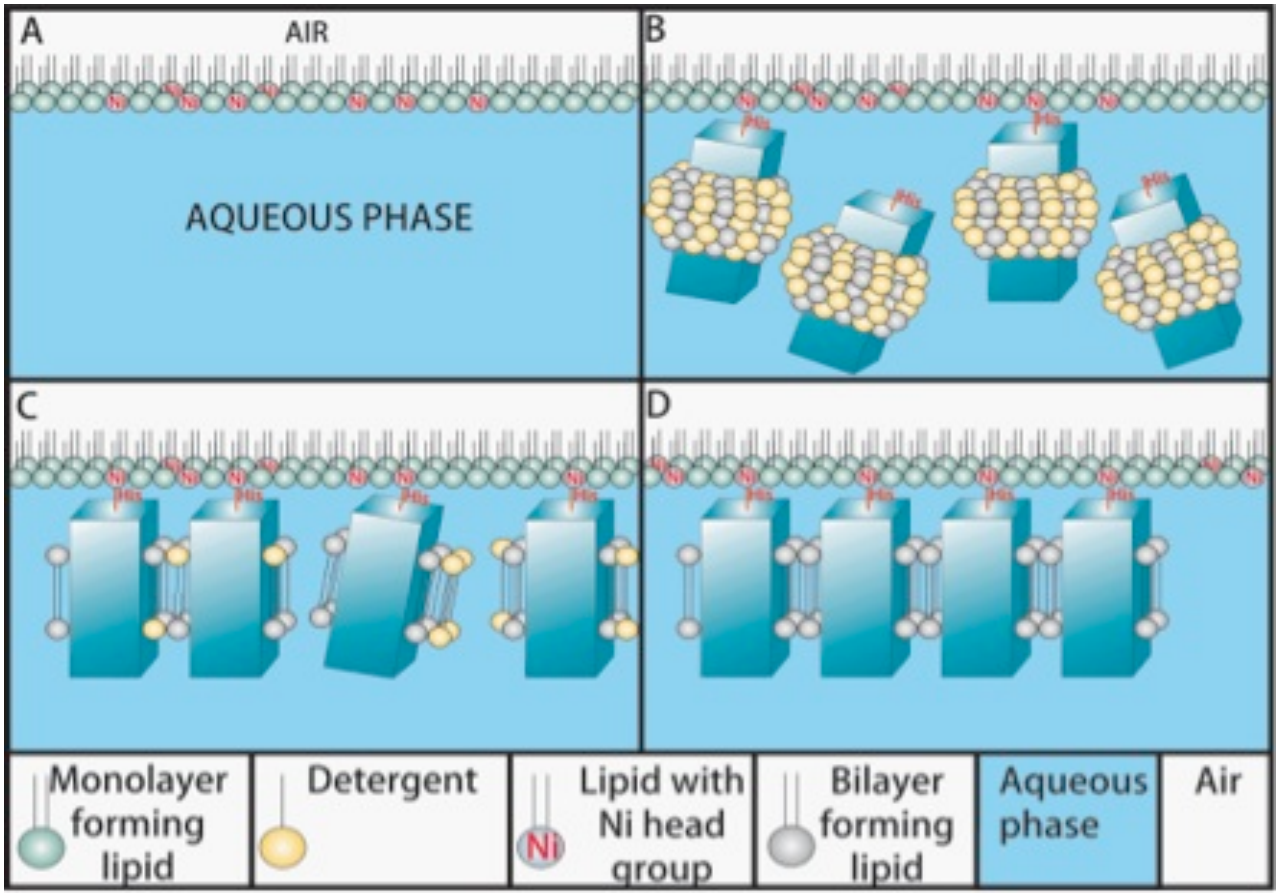


Figure 1.20 Illustration of Monolayer crystallization.

(Image courtesy of Dr. Ben Hankamer)

1.4.4.2.2 Monolayer crystallization

Monolayer crystallication (Lebeau and Venien-Bryan 2013) is another method which can help 2D crystals of membrane and soluble proteins for high-resolution electron crystallography. It involves the use a flat lipid monolayer (Figure 1.20A) as a template to produce 2D crystals of membrane proteins. In the example shown in Figure 1.20A, the monolayer is doped with functionalized Ni-chelating lipids (Figure 1. 20A) for the crystallization of His-tagged proteins, respectively. The His-

tagged protein is then absorbed onto the lipid monolayer (protein binding) (Figure 1. 20B), which forms a single layer closely packed with protein molecules, again initiating the crystallization process in an ideal scenario (Figure 1. 20C). For membrane proteins, an additional bilayer forming lipid is mixed with the detergent-solubilised, His-tagged protein, and detergent removed by the addition of detergent-binding beads (Bio-Beads) under the monolayer.

After some translation and rotation of the closely packed protein, 2D crystal patches are typically detected under favourable conditions within 1-2 days (Figure 1. 20D). Crystallisation of soluble proteins (e.g. Ferritin (Figure 1.21)) can result from the localized concentration effect that results from proteins being specifically bound at the monolayer surface.

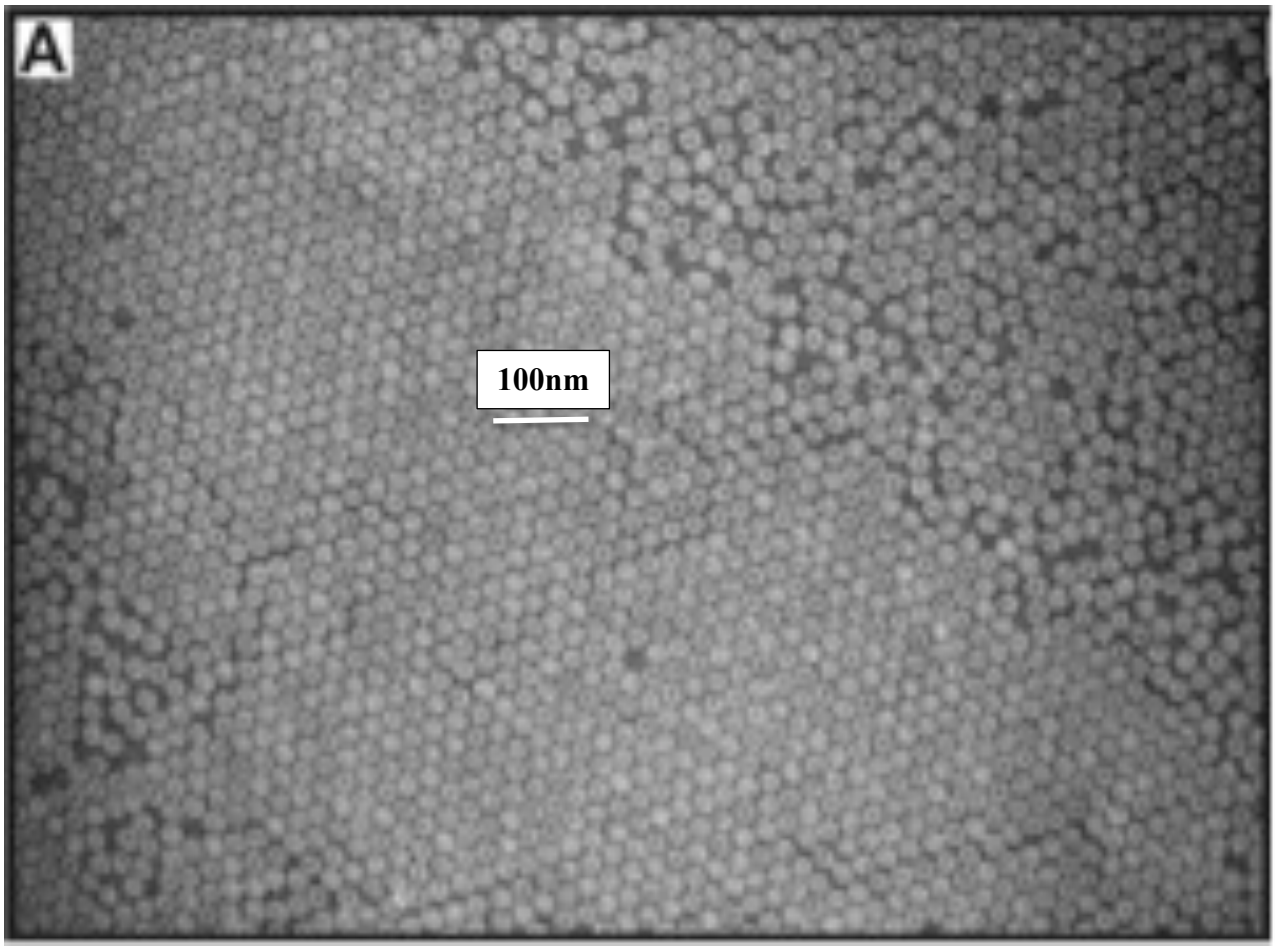


Figure 1.21 Example of production of 2D Ferritin crystals.

Ferritin particles bound to the monolayer are able to move along its surface to form 2D crystals. Image courtesy of Dr. Michael Landsberg.

1.4.4.3 Electron tomography (ET)

Electron tomography (ET) is a technique used to reconstruct subcellular and macromolecular structures in 3D from a tilted 2D series of projection images acquired using a transmission electron microscope. It is particularly well suited to studying structures, which cannot be averaged together (e.g. irregular cell structures or flexible macromolecular assemblies) and has thus become a

powerful tool for investigating the relationship between 3D structure and function of biological. ET provides structural details that can facilitate the mapping of high-resolution protein structures within a subcellular context and thus bridges the critical scale gap (Figure 1.22) between live-cell imaging and atomic resolution structure determination (Koster, Grimm et al. 1997, Trepout, Taveau et al. 2010).

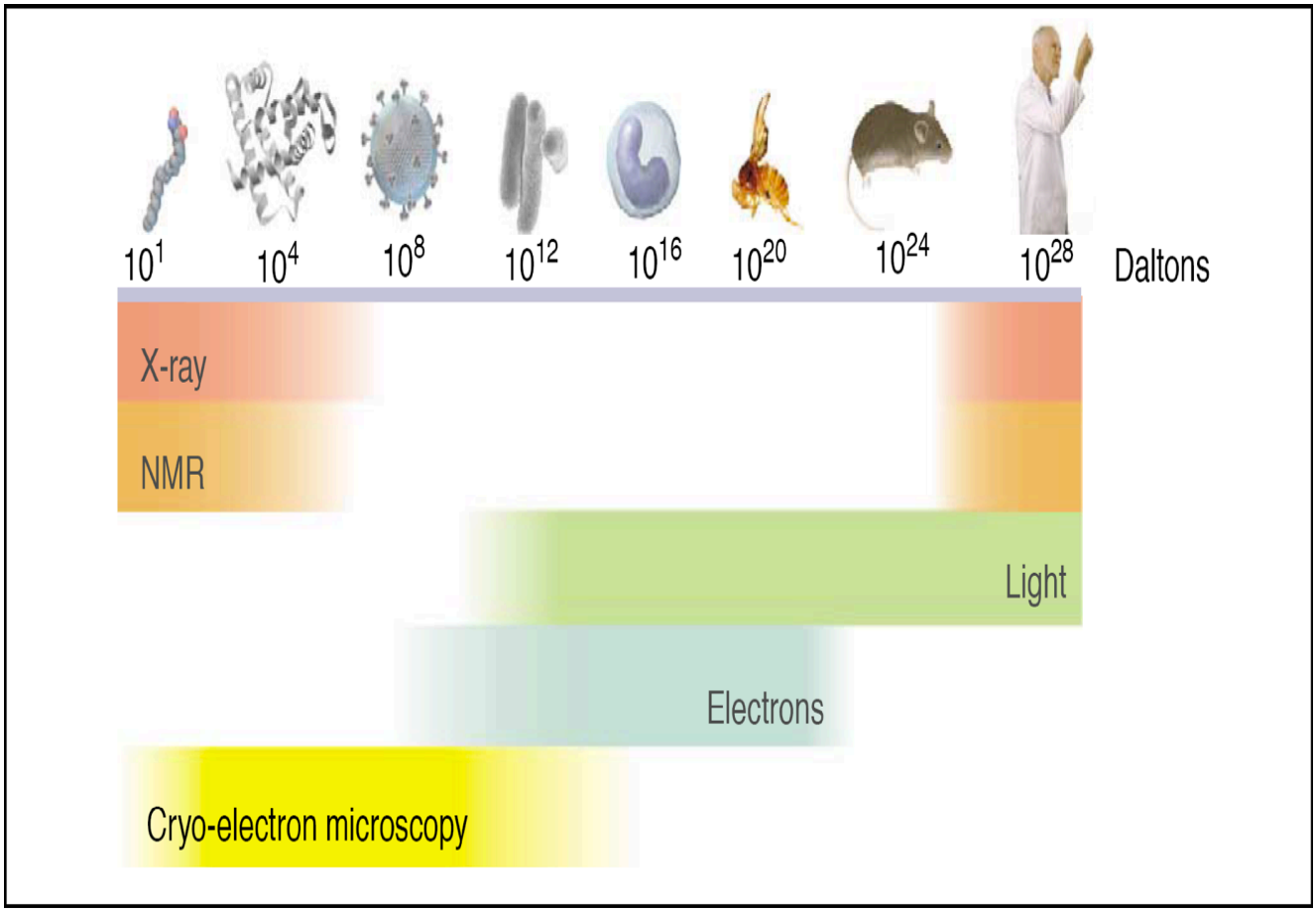


Figure 1.22 Resolution range achieved using various imaging techniques.

The size of objects is also shown which can be imaged in 3D with various techniques (Subramaniam 2005).

The best-known application of the principles of tomography is in CAT/CT scans used in medical/clinical imaging, in which an X-ray beam and detector rotates 360° around the patient to collect a series of projections images taken at incremental angles. This scanning yields a 3D image volume displaying the patient's internal organs at different densities. In cellular tomography, an electron beam is passed through the center of a thin (<300 nm) cellular specimen at incremental degrees of rotation. Projection images (tilt series) at different tilts (1° - 2°) are recorded using TEM and the tilt series is aligned over the angular range (Gan and Jensen 2012).

An important point for cellular electron tomography is that the range of tilt angles is limited to ±60° or ± 70° around one or two tilt axes. This is due to the thickness of the non-transparent sample holder required to hold the specimen in place, which at higher tilts obscures the electron beam. This

loss of information results in distortion and non-isotropic resolution in 3D reconstructions, in the direction of the beam (Nickell, Kofler et al. 2006) (Liu, Wright et al. 2010) as a direct consequence of the ‘‘missing wedge” of information (for a single axis tilt series) (Shaikh, Gao et al. 2008) for which Fourier coefficients cannot be obtained experimentally. The missing wedge can be reduced to a “missing pyramid” by collecting a second tilt series using a tilt axis orthogonal to the first (Figure 1.23).

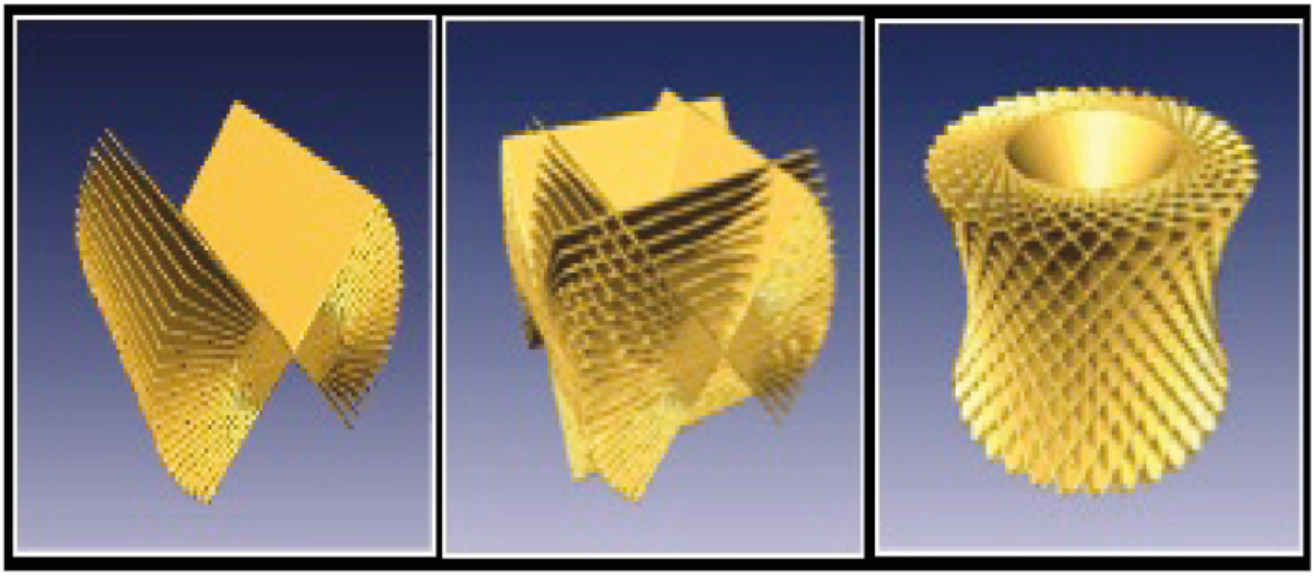


Figure 1.23 A comparison between single tilt, double tilt and conical tilt series.

The empty regions show the missing wedge in tilt. (A) The shape of missing region: a double wedge. (B) The shape of missing region: a double pyramid. (C) The shape of missing region: a double cone (Shaikh, Gao et al. 2008).

A recent research study (Paavolainen, Acar et al. 2014) introduced the sMAP-EM reconstruction method which is designed to compensate for the missing wedge artefacts. This method is reported to improve contrast ratio, significantly reduces the effect of missing information and produces isotropic resolution by recovering a proportion of the missing projection region. The method reported significant improvements in contrast and 3D resolution of reconstructed structures. More recently, (Moriya, Acar et al. 2015) demonstrated that the sMAP-EM method is capable of filling missing wedge and suppressing angular sampling artefacts during the initial estimation of 3D density maps. It has an additional advantage of quick visualisation of protein densities without requiring post processing.

The quality of a tomogram does not only depend on specimen preparation (plastic embedded vs frozen hydrated) and imaging techniques, but also on the processing algorithms used to recompute the 3D volume. In practice, after data is collected using EM, the next step is to reconstruct (Figure 1.24) the aligned tilt series into a tomographic volume.

There are many software packages available, which align (basically image rotation and translation) the projection tilt series and determine the tilt angle and tilt axis prior to 3D reconstruction. These packages include IMOD (Mastronarde 1997), EM3D (Ress, Harlow et al. 1999), TOM (Nickell, Forster et al. 2005), TxBR (Lawrence, Bouwer et al. 2006), protomo (Winkler 2007), TomoJ (Messaoudii, Boudier et al. 2007), SPIDER (Shaikh, Gao et al. 2008), BSOFT (Heymann, Cardone et al. 2008) and RAPTOR (Amat, Moussavi et al. 2008). Methods, like tomogram processing (reconstruction), noise reduction, segmentation, visualization, structure recognition, feature extraction, pattern recognition and interpretation need to be automated as much as possible.

Tomograms are stored and viewed as a series of 2D slices, stacked along the Z axis. In order to decompose the tomogram into its structural parts and obtain useful structural and biological information from 3D complex data, tomograms need to be segmented. Exploring the shapes of an object and interpreting an electron tomogram requires decomposition of the tomogram into its structural parts. Edge detection is considerably important particularly for cell and structural biology because of its extensive demand in applications such as particle picking and the annotation of sub cellular topographic volumes, in which edges define boundaries and the regions of interest lies between these boundaries. A combination of high-resolution segmentation techniques allow us to localize specific molecules *in situ* in 3D and provide the temporal and spatial coordinates to detect tiny objects, but the performance of segmentation techniques is highly dependent on denoising approaches. Preprocessing of the volumes is the first step in discerning the molecular components of an electron tomogram, This preprocessing typically involves denoising and removal of artifacts introduced by CCD and EM. Therefore, interactive and automated noise reduction tools clearly play a vital role in obtaining high-resolution structures of macromolecules by electron tomography. The removal of noise by denoising algorithms can greatly help both manual and in particular partly or fully automated segmentation.

1.4.4.4 The influence of noise in Electron Tomography

The ability of a method to accurately detect edges within electron tomographic data is paramount for the segmentation of tomographic data, but this ability is often undesirably affected by spurious variations in brightness and contrast called noise. However, edges that describe real biological features (e.g. organelles and macromolecular assemblies) are also simply significant changes in voxel intensity. Complicating this further is that biological samples are very sensitive to beam/radiation damage and therefore, low dose imaging conditions are generally employed for the acquisition of cellular electron tomography data, resulting in noisy tomograms. To improve contrast, cells are usually stained with heavy metal salts (e.g. uranyl acetate, lead) during preparation, which enhances interactions between incident electrons and stained membranes during

imaging. However, like with single particle methods, heavy metal stained specimens suffer from reduced resolution compared to cryo-tomographic analyses. In order to obtain improved reconstruction quality, tomographic volumes are often subjected to a de-noising prior to segmentation analysis, with the aim of reducing noise while preserving edge information.

A wide range of the denoising filters has been reported in the literature and the performance of these is summarised in the sections that follow. The most important issues that need to be considered in comparing filters are interactive parametric tweaking, memory management and time management (Sandberg 2007). An ideal filter offers minimal demands on time and memory resources and little or no subjective parameter setting.

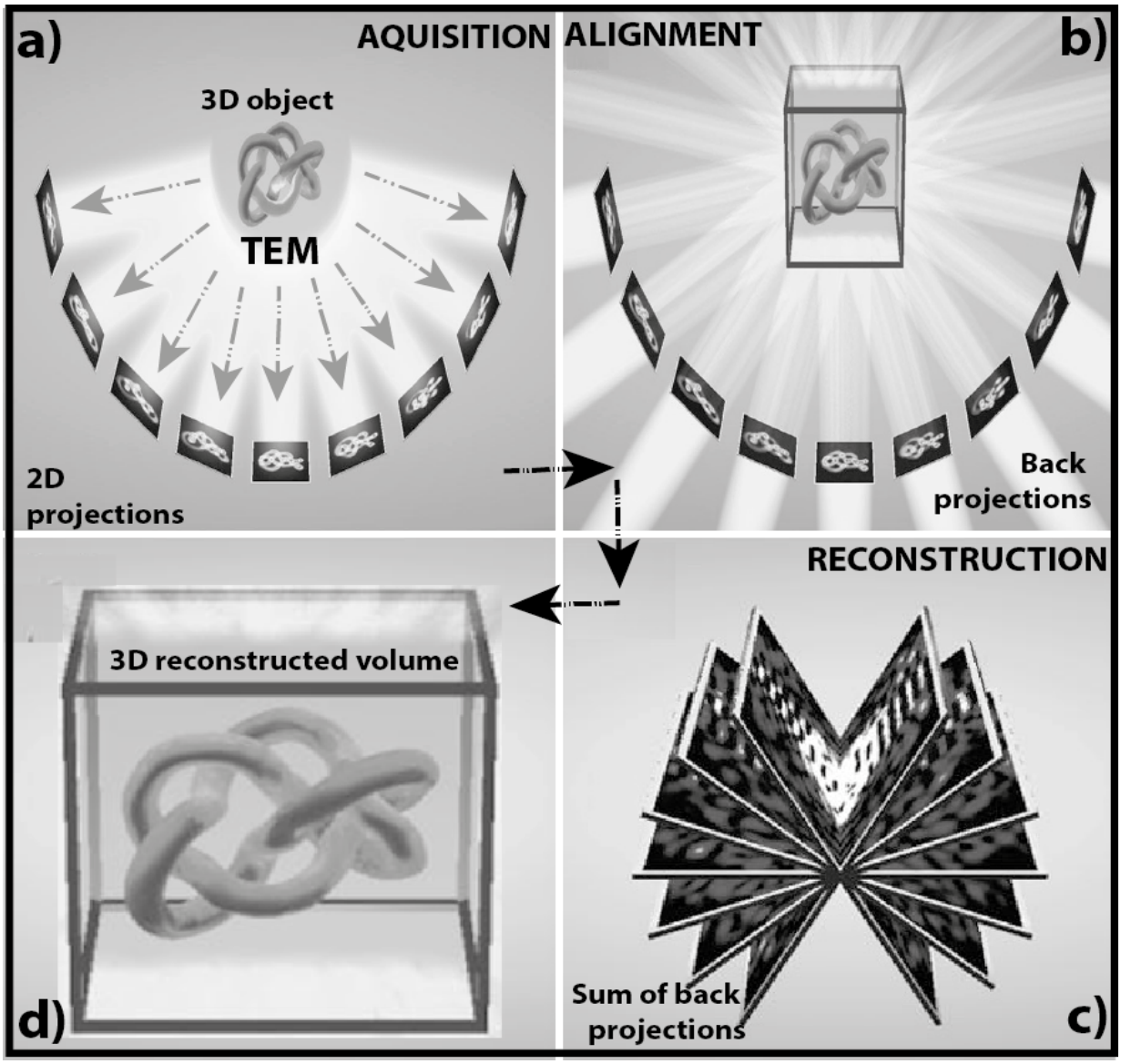


Figure 1.24 Schematic representation of process of tomographic reconstruction.

Images show the acquisition, alignment and reconstruction schemes for tomographic reconstruction, in which (a) a specimen is placed under TEM, and either specimen or detector was tilted incrementally to record a series of 2D projections. (b) For each 2D projection, back projection body is calculated, aligned. (c) All the back projections are then summed up all the back projection in Fourier space to reconstruct the 3D

volume. (d) The final 3D reconstruction (tomogram) of input object is generated. The image was adopted from (Baumeister, Grimm et al. 1999).

1.5 Computational methods for image processing of ET data

A range of image de-noising, edge detection and segmentation software is being developed to automate the segmentation of organelles and macromolecular assemblies. These are designed to facilitate the extraction, processing and annotation of sub-volumes as well as the docking of high-resolution structures into them on the path to developing atomic resolution 3D atlases of cells. In the section below noise reduction filters, edge detection and segmentation techniques are briefly summarized.

1.5.1 Noise reduction filters

1.5.1.1 Fast Fourier Transformation (FFT) based filters

In image processing, Fourier transformation (FT) is a tool that decomposes an image signal into a set of frequencies representing the original image. FT decomposes an image into its sine (imaginary) and cosine (real) components, which represent the amplitude (Figure 1.25A), and phase shift (Figure 1.25B) of the image respectively, and in doing so it converts the image from the spatial domain to the frequency domain. Due to image distortion during data capture, image processing often involves the removal of high frequency and low frequency noise.

Fast Fourier Transformation (FFT) filtering techniques (such as low pass, high pass, and band pass) are the simplest and most effective noise reduction techniques. FFT filtering techniques are capable of removing undesired spatial frequencies including high frequency (impulse) and low-frequency (e.g. stain gradients) noise (Lucic, Forster et al. 2005). Through the selection of appropriate 'bandpass' settings, FFT filters can be used to select the frequency range of interest. However, by excluding specific frequency ranges indiscriminately, both the signal and noise outside of the bandpass range are lost.

For instance, FT followed by a 'high pass' filter is frequently used to remove unwanted low-resolution data from a microscopic image in TEM image formation. Likewise 'low pass' filtering is used to remove 'high-frequency' data from the image. On the selected frequency domain set, inverse Fourier transformation can be performed to recover a filtered image in its original spatial domain.

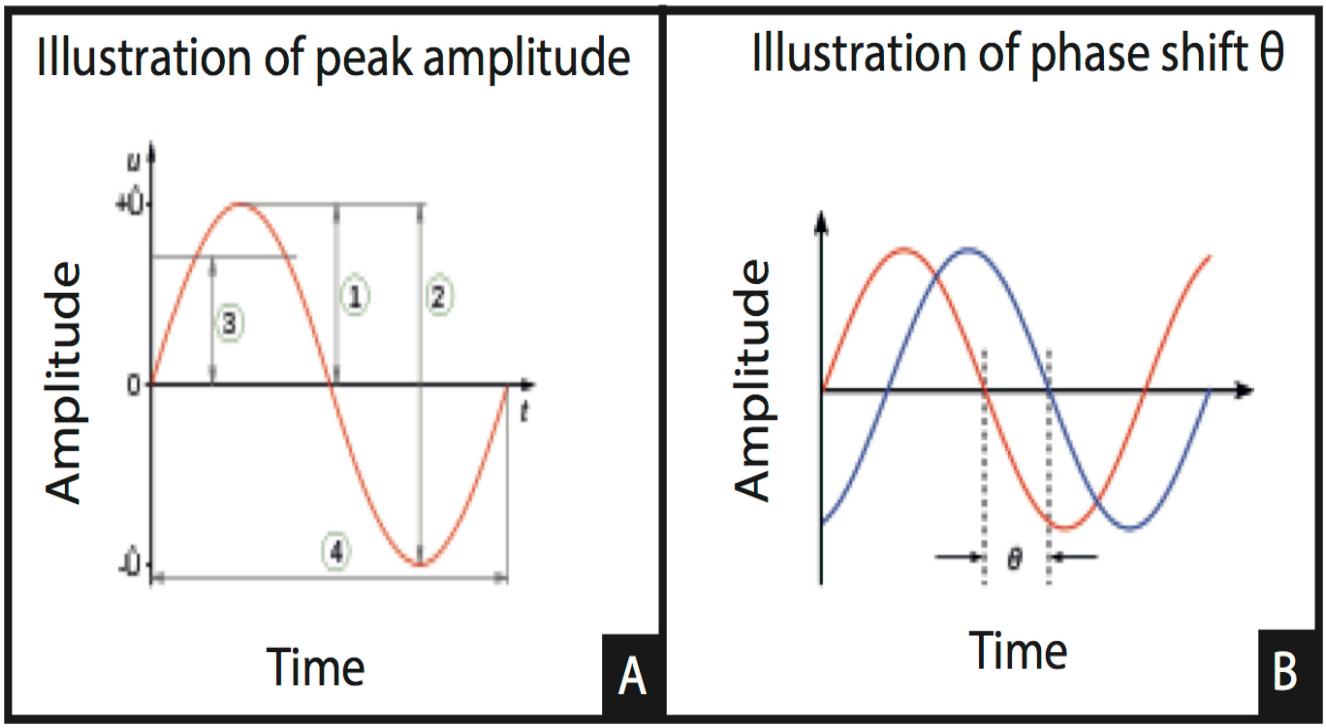


Figure 1.25 Illustration of peak amplitude and phase shift (θ).

1.5.1.2 Kernel based filtering schemes

An alternative approach is to use *kernel filters* such as Gaussian, mean or median filters. In these approaches, a kernel is designed to conduct specific operations. For example the mean filter replaces each pixel value with the mean of all the neighboring pixel values within the kernel window and yields smoothed images. In so doing the mean filter suppresses pixel values, which are unrepresentative of their neighborhood (e.g. impulse noise) and reduces the amount of intensity variation between pixels. Gaussian kernels (Pantelic, Rothnagel et al. 2006) calculate weighted averages based on pixels within the radius of the Gaussian filter. This filter replaces the current value of a focal pixel with the weighted sum of pixels in its surrounding space to smooth the image and attenuate noise. Pixels directly adjacent to the focal pixel have a greater influence on the weighted average than those towards the outer edge of the Gaussian radius. Due to its suitability to biological applications, computational biologists have developed many extensions of this filter. These include non-linear Gaussian, Gaussian notch, Gaussian-median, difference of Gaussian and Laplacian of Gaussian filters (Nagayoshi, Murase et al. 2005, Catarious, Baydush et al. 2006, Wei, Wang et al. 2012, Kong, Akakin et al. 2013).

1.5.1.3 Wavelet transformations

Wavelet transformation for noise reduction which has been suggested to outperform routinely used filters, such as median filter and low-pass (Stoschek and Hegerl 1997). The approach is most suitable for images with high spatial frequency information as it preserves high-resolution

information relatively well based on the fact that the image signal is decomposed into blocks of data that are transformed into wavelet domains. However, the wavelet transformation has a high computational cost and for many real 3D applications such as tomogram de-noising is almost unaffordable purely from a computational perspective.

1.5.1.4 Nonlinear anisotropic diffusion filter

Nonlinear anisotropic diffusion (NAD) is a simple, fast and efficient noise reduction algorithm with good signal preservation (Frangakis and Hegerl 2001, Yamashita, Zhang et al. 2007, Davies, Anselmi et al. 2012). It is a real space hybrid approach, which is based on two diffusion fluxes:
Edge enhancing diffusion (EED), which effectively preserves and enhances edges. It inhibits smoothing across edge structures while still diffusing along them. The edges are preserved and enhanced by favoring inter-region diffusion over intra-regions. Coherence enhancing diffusion (CED), which focuses on disconnected lines and enhances them according to the strength and direction of the smoothing by means of a structure tensor. It was originally designed to enhance friction ridge structures i.e. thumb impressions.
NAD filters are generally regarded as better de-noising approach than low pass, median and wavelet transform filtering due to their ability to apply adaptive diffusion filtering near edges to preserve them. Diffusion is a process that equilibrates concentration differences between similar pixels (either edges or noise) without creating or destroying pixel intensities or the gradient magnitudes. To perform diffusion steps, a Gaussian function is adaptively used to make the process non-linear and to impose a blur on the original image and noise. NAD then replaces the grey values of an image with the resultant diffused values until the target (user supplied rate of diffusion threshold) is reached without signal degradation (approximately 50-60 iterations). In particular, NAD can easily enhance features such as membranes at molecular resolution (Lucic, Forster et al. 2005, Yamashita, Zhang et al. 2007). This method has yielded useful information about complex biological structures while reducing the noise efficiently and is considered an excellent signal preservation technique. However like other diffusion-based techniques, it is time consuming and requires a proper adjustment of at least two parameters. For example, the 3D median filter requires computer memory to be allocated equivalent to approximately twice the number of Z slices (i.e. 2 x the number of voxels being considered) in the window. In comparison, NAD filtering requires around 36 times the total number of voxels in the volume (van der Heide, Xu et al. 2007). NAD filtering is incorporated as a standard feature of the IMOD software package (Alber, Forster et al. 2008) and Bernard's Software Package (BSOFT) (Frangakis and Forster 2004, Heymann and Belnap 2007).

1.5.1.5 Iterative median filters

The median filter is a robust, non-linear noise reduction technique. It preserves edges, which tend to get blurred by linear filters (such as mean, low pass) under certain conditions. The main problem in the application of the median filter is degradation of high frequency values because it scans through the whole volume and indiscriminately replaces all intensity values with the median (taken from a ranked list of values) without considering whether a pixel is a true or false positive. Thus the application of this filter is often subjected to a subsequent edge-sharpening step.

The iterative median filter is an efficient, automatic noise reduction protocol for 3D ET reconstructions of biological materials that addresses some of the shortcomings of the standard median filter (van der Heide, Xu et al. 2007). The iterative median filter lets a 3D kernel replace the voxel values with the median value. Iteration of linear filters results in edge blurring but the iterated median filter will eventually reach a stationary "fixed point" and can enhance the SNR as signal and noise reach clearly different median values after a finite number of iterations. Although, the iteration process decreases the variance, it has the advantage that it preserves edges. To remove high levels of noise while preserving high frequency information, it is often better to use a large number of iterations rather than increase the kernel size of the median filter. In one particular study, three iterations of median filtering was shown to satisfactorily enhance SNR in most cases, as this provided the largest gain in noise reduction performance relative to computational cost. Beyond this, improvement in the noise reduction increasingly diminished (van der Heide, Xu et al. 2007).

1.5.1.6 Bilateral Filter

The bilateral filter is a real space anisotropic (directionally dependent) filter based on a non-iterative, edge-preserving smoothing function (Figure 1.26) capable of detecting the discontinuity of edges. The bilateral filter was first designed to denoise both RGB photographs and gray-scale images (Tomasi and R.Manduchi 1998), but was later demonstrated to be a useful tool for a wide range of EM applications including ET and SPA (Jiang, Baker et al. 2003). It replaces the original focal pixel with the bilaterally weighted sum of neighbouring pixels in its immediate area (Jiang, Baker et al. 2003, Pantelic, Rothnagel et al. 2006). Bilateral weights (Figure 1.26) are calculated such that pixels on the same side of an edge contribute to the new focal pixel value (Jiang, Baker et al. 2003, Pantelic, Rothnagel et al. 2006). Thus, the bilateral filter combines spatial domain and range (intensity domain) filtering.

The main disadvantages of the bilateral filter are that it is unable to distinguish between impulse noise and true edges and requires adjustment of two parameters. Because images taken at different

defoci contain have different contrast and it can be hard to predictably optimize filtering parameters.

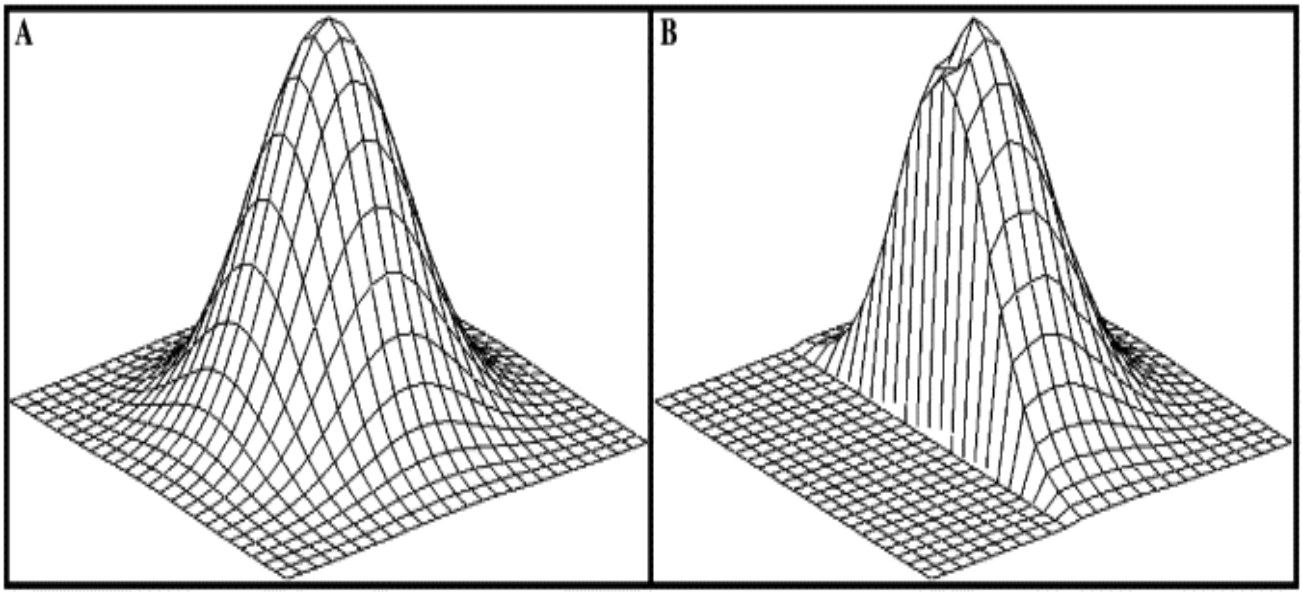


Figure 1.26 The Gaussian and bilateral filters.

The focal pixel is located under the peak of the Gaussian function in both A and B. (A) Gaussian filter. The mesh represents the shape of the Gaussian function. When Gaussian filtering is applied a new intensity is calculated for the focal pixel based on the weighted average of intensities of neighbouring pixels under the mesh. It assigns more weight to the focal pixel and less weight to the neighbours. (B) Bilateral filter. The mesh describes the shape of the bilateral filter. In a local region containing an edge, the bilateral filter replaces the focal pixel's intensity by the weighted sum of intensities of neighbouring pixels under the mesh and on the same side of the edge e.g. right side of the mesh in this figure, which means the pixels to the left side of the edge are excluded from the filtering process. The bilateral filter therefore smoothens noise and preserves edge detail. Figure adapted from (Pantelic, Rothnagel et al. 2006).

1.5.1.7 Discriminative bilateral filter (DBL):

The DBL is an extension of the bilateral filter and is particularly suited to low SNR and high impulse noise images (Pantelic, Rothnagel et al. 2006). The DBL filter efficiently classifies impulse noise pixels and edge-forming pixels by observing their inherently different properties. The DBL filter replaces the original focal pixel with a discriminative bilaterally-weighted sum of neighbouring pixels. The filter behaves like a standard bilateral filter when there is no noise pixel, but when a pixel is identified as impulse noise in the filtered region, it operates in two different ways. First, when impulse noise pixels are present in the filter region (but are not the focal pixel) the DBL detects such pixels and excludes their contribution from the smoothing process. Second, when the focal pixel itself is corrupted by impulse noise, the DBL will use the local median intensity (within a radius of 1 from the focal pixel) in place of the original focal pixel intensity in calculation of bilateral weights and summations.

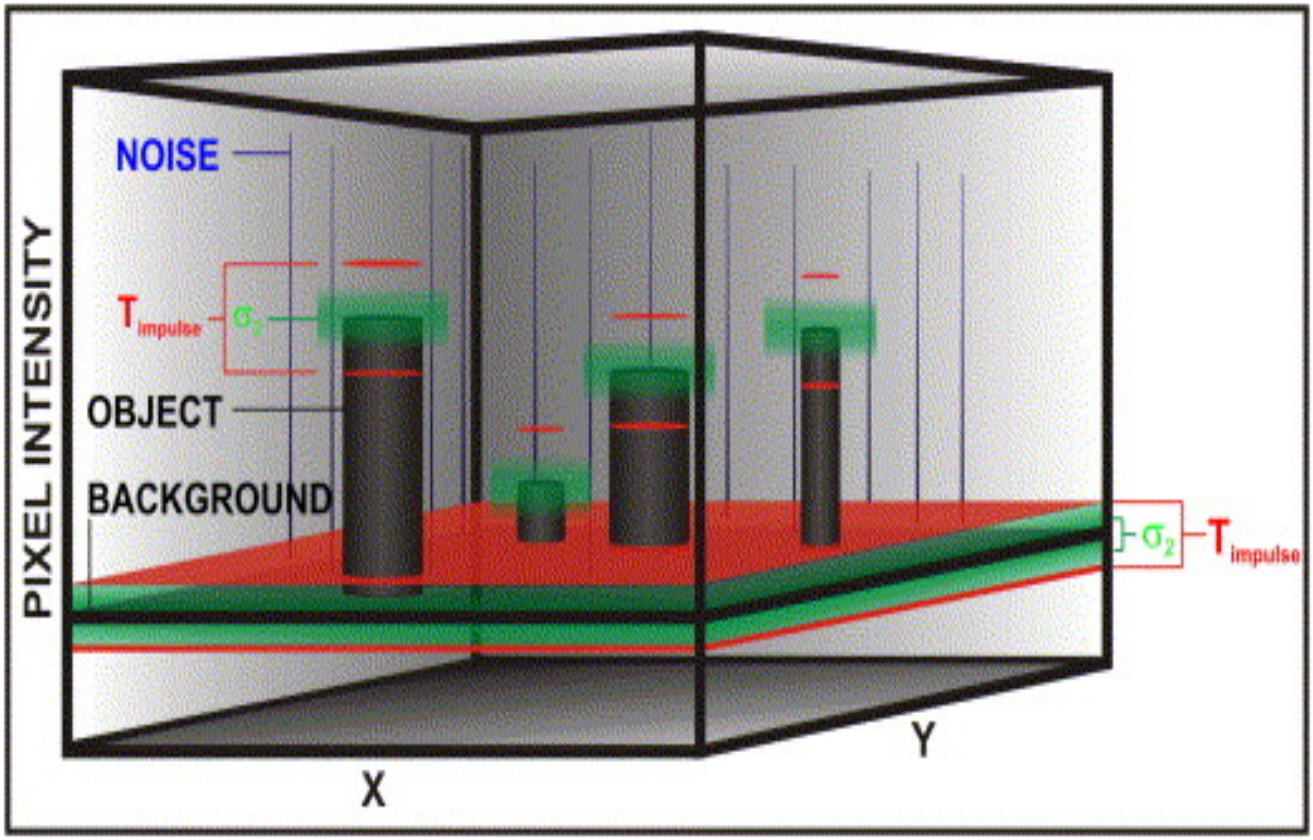


Figure 1.27 The DBL filter concept.

This diagram is a 3D representation of a 2D image. X and Y axes represent X and Y dimension of the image, whereas Z axis represent pixel intensities. A horizontal black bar represents background pixel intensities. Four black vertical columns coming out of the background intensities represent signals (objects) whereas the blue vertical lines are impulse noise spikes. The DBL filter scans through the whole image pixel by pixel and resulting pixels are based on the parametric settings of σ1, σ_2 and $T_{impulse,}$ where σ_1 defines the radius of the filter and set to $3\sigma_1$. σ_2 is a photometric parameter of the DBL. It effectively defines the difference in intensity values. The green band depicts the intensity range (Z axis). $T_{impulse}$ is a local threshold parameter to discriminate between edges and impulse noise. The red bars depict this parameter. To a first approximation, this threshold defines a pixel as impulse noise if the difference between intensity values of focal pixel and neighbouring pixels is more than $T_{impulse.}$ Figure adapted from (**Pantelic, Rothnagel et al. 2006**).

One of the major advantages of the DBL filter is the reduction of high frequency noise while preserving genuine high frequency information as is seen in Figure 1.28. As such, this filter may be more effective in filtering smaller structures in biological applications with low SNRs, as noise suppression is less affected by filter size. The DBL filter should thus provide more focused smoothing than either the Gaussian or bilateral filters.

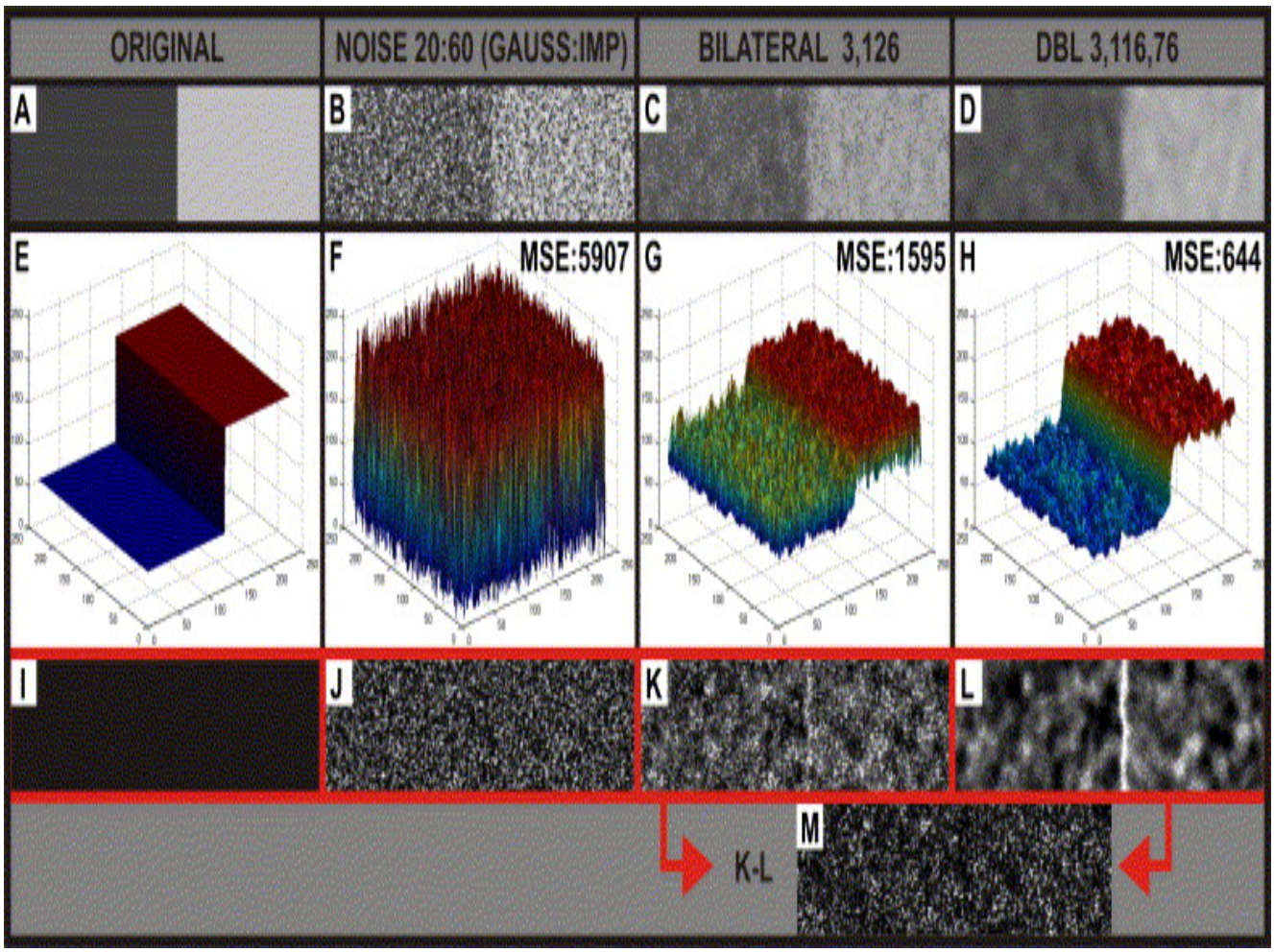


Figure 1.28 Evaluation of DBL and bilateral filter performance.

(A) Original grey-scale image. (B) Original image (shown in A) contaminated with 20% Gaussian: 60% impulse noise. (C) Application of bilateral filter on noise contaminated image (shown in B). (D) Application of DBL filter on noise contaminated image (shown in B). the DBL filter functions as a bilateral filter at $T_{impulse}$ = 255,. (E–H) Colour contour plots of A–D. Both the bilateral (G: MSE = 1595) and DBL (H: MSE = 644) filtered images show a marked improvement as compare to noise contaminated image (F: MSE = 5907). (I) is the control (i.e., A minus A) for auto-equalized difference testing. (J–L) show auto-equalized difference between original, noise contaminated, and filtered counter parts (A, B, C and D respectively) e.g. (J) is the difference image between (B) and (A). The noise level is considerably lower in the DBL filtered image. (M) The difference image (K minus L). Figure adapted from (Pantelic, Rothnagel et al. 2006).

1.5.2 Evaluation of noise reduction techniques

Noise reduction and edge detection techniques are important for structural analysis of objects in the field of electron tomography as well as in the broader bio-imaging field. Each technique described above has its own advantage and disadvantage. The filtering performance is dependent on a number of parameters, computation time and memory requirements. Wavelet transform is computationally expensive however it enhances the performance of lossless image compression (Reichel, Menegaz et al. 2001). The NAD filter can provide faster performance than the wavelet transform technique (Yamashita, Zhang et al. 2007, Volkmann 2010), but requires great care and the adjustment of 4 parameters to achieve optimal results. It fails to preserve high frequency information, if the 'Edge

enhancing diffusion (EED)' and 'Coherence enhancing diffusion (CED)' parameters are not chosen carefully. It can preserve edges at low SNR, but is also extremely computationally expensive, as it requires around 36 bytes of memory for each tomogram voxel. In contrast the 3D iterative median filter is an automatic filter and requires computer memory to be allocated equivalent to approximately twice as many Z slices as the size of the window without tuning any parameters (van der Heide, Xu et al. 2007). The 3D iterative median filter suppresses noise efficiently in a non-linear fashion, and preserves edges, which tend to get blurred by linear filters. On the other hand, it degrades high frequency information and replaces all intensity values with an indiscriminate median. In comparison, the bilateral filter has only two adjustable parameters. It was adapted as a tool for ET and SPA (Jiang, Baker et al. 2003) and preserves edges as well as edge discontinuities with moderate accuracy. Despite the significant benefits of the bilateral filter, it remains unable to distinguish between impulse noise and true edges particularly when impulse noise levels are high.

The discriminative bilateral filter (DBL) was specially designed to account for impulse noise through an additional parameter (a total of three parameters) to be tuned. The DBL filter reduces high-frequency noise but preserves the genuine high frequency information however it is computational expensive (Pantelic, Rothnagel et al. 2006).

Filter Name (No. of Parameters)	**Advantages**	**Disadvantages**
Wavelet transformation	Enhances the performance of lossless image compression	Computationally expensive
NAD (4)	Faster than wavelet transformation technique	Computationally expensive, failed to preserve high frequency information.
Iterative median filter (Auto)	Efficiently working in non-linear fashion, preserve edges	Degrades high frequencies, replaces intensities with indiscriminate median.
Bilateral Filter (2)	Detects discontinuity in edges with moderate accuracy.	Unable to distinguish between impulse noise and true noise.
DBL (3)	Specially designed to account for impulse noise.	Computational expensive

Table 1.2 Evaluation of noise reduction techniques.

Contrast loss during image formation in single particle electron cryo-microscopy hinders identification and interpretation of high-resolution structural features. To overcome this issue (Fernandez, Luque et al. 2008) developed a method which reportedly helps to recover high

resolution molecular features by automatically determining the B-factor of high resolution maps. In particular it aims to restore the contrast of a density map without noise amplification. The TOMOBFLOW (Fernandez 2009, Fernandez and Carrascosa 2010) is another nonlinear de-noising technique based on 'Beltrami flow'. This noise-reducing method reportedly efficiently preserves features and retrieves biologically relevant information.

In summary, based on information above, the iterative median filter can be easily used with the added advantage of no adjustable parameter, less time complexity and less memory requirements. The median filter is reported to work well for cellular tomograms (Sandberg 2007) because of its simplicity. The bilateral filter has the advantage of preserving edges better while the DBL filter additionally suppresses impulse noise.

1.5.3 Edge detection and segmentation in electron tomography

Edge detection in raw and de-noised images is of considerable importance for applications such as single particle picking, the annotation of sub cellular tomographic volumes and to provide boundaries for molecular docking. Tomograms are stored and viewed as a series of 2D slices, stacked along the Z axis. In order to decompose the tomogram into its structural parts and obtain useful 3D structural and biological information from complex 3D tomographic datasets these must be segmented. By using a combination of de-noising and segmentation techniques it is possible to localize not only membranes but also macromolecular assemblies within the 3D volumes of cells thereby identifying their spatial coordinates and yielding contours for annotation and molecular docking. The performance of segmentation techniques varies depending on the size and shape of different biological structures such as thylakoid membranes and ribosomes. Many edge detection techniques have been developed, but are generally tailored to specific applications (Gonzalez 2002). With high SNR data sets, these techniques may be adequate, but these simple edge detection techniques generally do not deal well with complex, noisy data sets containing numerous overlapping objects and poorly defined edges, and do not extend well into 3D. Methods based on intensity and contrast traditionally have problems in detecting thin, elongated structures, for example, but these can be determined from geometric parameters (Sandberg 2007).

There are many challenges faced by segmentation methods specific to cellular tomograms. One of them is a great variety in datasets because of different specimen preparation and imaging methods as well as diverse biological structures. Contrast is usually non-uniform and low along membranes and tubular structures. The signal intensity of membranes is also typically lower than the contrast of surrounding ribosomes and other interfering structures making them more difficult to detect. Point-like objects such as ribosomes usually have a strong uniform contrast, but they overlap with the membranes complicating the segmentation of the latter. Another issue is low SNR. The "missing

wedge" typical of TEM tomography introduces the problem of anisotropic resolution. Some areas of the tomogram will also have lower contrast than others. While dual axis tilting (Hohn, Tang et al. 2007) and particle averaging (Forster and Hegerl 2007) are able to increase SNR to an extent, the issue of non-uniform contrast across tomograms remains a considerable problem.

As with image filtering, excessive computational complexity and the need to set optimal input parameters can hinder automated segmentation methods. Segmentation efficiency not only depends on whether the method is manual, semi-automated or fully automated, but also upon the number of input parameters required (related to noise, scale level or numerical properties of the algorithm) and whether they are independent or require a combination of values to be set. To set three independent values optimally is much easier than identifying one optimal set of three different values, for example. 3D segmentation algorithms are also generally computationally intensive. An ideal 3D segmentation algorithm should therefore consider processor and memory requirements. The primary objectives of such an algorithm should be, accuracy, speed, and limited dependence on user-defined parameters.

1.5.3.1 Manual Segmentation

Manual segmentation is very time consuming and requires biological expertise. For example, to manually trace the regions of interest for the 3.1 x 3.2 x 1.2 μm^3 cell volume presented in (Marsh, Mastronarde et al. 2001) required around 9-12 months using IMOD (Hohn, Tang et al. 2007). In addition to this substantial time requirement, the impacts of human error, although often very slight, can be significant.

1.5.3.2 Semi automated approaches

To overcome these problems a number of automated and semi-automated segmentation techniques have been developed. Tools such as the 'drawing tool' and 'interpolator tool' plug-ins (Noske, Costin et al. 2008) have extended the manual tracing capabilities of the IMOD software package (Alber, Forster et al. 2008). Using the "drawing tool" allows direct warping and sculpting of pre-drawn, simplified contours, while Noske's "interpolator" plug-in interpolates the contours drawn between two different slices. For the later, reasonable results can be obtained by tracing every 5^{th} slice manually within a 3D stack, following which interpolator automatically generates contours for the intermittent slices and to generate a final 3D surface-rendered model. Figure 1.3 demonstrates the segmented compartments of an algal cell, generated using semi-automated tools (Noske, Costin et al. 2008). Each compartment is marked up in a different colour for identification purposes.

1.5.3.3 Thresholding

The simplest approach to segmentation is thresholding, which converts a grayscale image to a binary image by applying an optimal threshold (Shapiro and Linda 2002). It is extended into 3D through isosurface contouring. Processing is based on the image's gray values by defining a range, outside of which values are discarded. In standard thresholding, all the intensity values lower than the threshold will be replaced with a grayscale value of 0 (black), and the rest with a grayscale value of 255 (white). Hysteresis thresholding uses two threshold values (high and low) which typically improves connectivity within edges (Canny 1986). Any pixel above the high threshold is turned white and considered to be a strong edge pixel. Hysteresis filters then recursively search the surrounding pixels. If the values are greater than the low threshold they are also turned white and retained, but if these values are not connected to a high threshold value the edges are removed.

1.5.3.4 Gradient based edge detectors

In image processing, the Prewitt operator (Judith M. S. Prewitt 1970), Roberts cross edge detector (Roberts 1963) and Sobel edge detectors (Gonzalez 2002) are considered classical edge detectors based on image gradients. The Prewitt operator computes an approximation of a gradient of intensities and detects directional (horizontal and vertical) edges by calculating the difference between pixel intensities in respective directions. The Roberts cross edge detector computes the gradient by calculating the sum of squares of the difference between diagonally adjacent pixels in an image. This edge detector has a simple kernel and produces only integers. Similarly, Sobel edge detectors work on the same gradient principle but emphasize high spatial frequencies, which correspond to edges by calculating the rate of change of the focal pixel with respect to its neighbouring pixels. These classical filters are computationally inexpensive, but very sensitive to noise. High levels of noise may result in degradation of edges and loss of quality.

1.5.3.5 Canny edge detector:

The Canny edge detector is a multi-stage complex filter developed for edge detection (Canny 1986). To optimize the detection probability of real edges and accurately locate detected edges, this algorithm sets it background based on three parameters. The first parameter sets the width of the Gaussian function, which controls the degree of image smoothing and the suppression of noise. It then, approximates gradients at each pixel position by using a Sobel operator (Gonzalez 2002). Edges are found depending on their magnitude (magnitude represents edge strength) and direction. Other parameters set threshold values for strong and weak edges. In this way, it detects true weak edges by preserving all local maxima to trace the edges, and suppressing everything else. To maximise the performance of the filter, the Canny filter is focused on detecting real but low SNR

edges (i.e. it aims to detect the maximum number of real edges), localization (i.e. the proximity of a detected edge to its original position) and minimal response (i.e. no false edge detection due to noise). The accuracy of the Canny edge detector is such that it is widely acknowledged as one of the best edge detection algorithms, but its complexity has hindered widespread use (Canny 1986). Additionally, the Canny filter fails to detect true discontinuity (i.e. to check whether the discontinuity is due to noise or a true edge), and therefore can produce spurious and unstable boundaries in highly noisy data (see Chapter 2).

1.5.3.6 The Snake algorithm

The Snake algorithm is an active contouring algorithm. The user manually draws a rough contour (snake) enclosing the structure of interest, and the snake interprets the contour semi-automatically by pushing the snake closer to the feature of interest (Kass, Witkin et al. 1988). The method is limited to finding one contour per 'guess'. It therefore does not work well with topological changes such as when one contour splits into two. With 3D data, the snake algorithm first locates all contours in one slice as an initial guess and then uses them for nearby slices.

1.5.3.7 The watershed transform

The watershed transform is a mathematical morphology-based segmentation technique, based on the concept of watershed and catchment basins (Roerdink and Meijster 2001). Geographically, a watershed is the ridge (geographical boundary) that divides areas drained by a river, whereas the catchment basin is the area draining into the river. The watershed transform is conceived from the original concept of watershed and applied to gray-scale image processing by considering the catchment basins as biological objects or regions to be segmented. As the catchment basins expand, watershed lines are placed (Volkmann 2002). The algorithm stops processing when the water level reaches its highest peak in the landscape.

Over segmentation is a known drawback of standard watershed algorithms. This can be suppressed by calculating the watershed following application of noise reduction algorithms (Sijbers, Scheunders et al. 1997). Another solution specifically designed for segmenting electron density maps is based on the appropriate selection of flooding step size for the advancement of the density level (Volkmann 2002).

1.5.3.8 Bilateral edge filter

The bilateral edge filter is another adaptation of the bilateral filter (Pantelic, Ericksson et al. 2007), and was specifically designed for the analysis of 2D electron micrographs. This technique is most suitable for images containing predominantly low- and mid-frequency information. It is a nonlinear,

photometrically weighted discontinuity-based anisotropic filtering technique. The aim of the filter is to suppress noise by attenuating undesired frequencies while enhancing edge detection by extracting specific features selectively. It identifies significant discontinuities, minimising the noise artifacts introduced throughout the extraction process. It has been shown to generate better results than the (LoG) (Gonzalez 2002), Marr-Hildreth (Marr and Hildreth 1980) and Canny edge filters (Canny 1986). The bilateral edge filter has the added advantage of requiring the adjustment of only one parameter. Furthermore, this filter has faster processing times when applied to standardized image sizes and can contribute to selective contrast enhancement. It preserves high-resolution details for particle picking and segmentation from low SNR SPA and electron tomography data. Its filter size (5 x 5 local neighbourhoods) is small and fixed which increases filtering speed. After noise suppression, the proper adjustment of its parameter settings can isolate strong/significant edges. The overall accuracy and discriminative potential of the bilateral edge filter makes it applicable to a wide range of functions including semi-automated segmentation of microscopic images, SPA and contrast transfer function (CTF) correction. It has already proven useful in particle picking as implemented in the semi-automated particle picking software $Swarm_{PS}$ (Woolford, Ericksson et al. 2007).

1.5.3.9 Laplacian of Gaussian & Arbitrary Z-crossings

The Marr-Hildreth filter (Marr and Hildreth 1980) convolves images with a Laplacian of Gaussian (LoG) function which involves a two-step process. In the first step it applies a Gaussian filter for pre-smoothing. In the second step it calculates the second derivative of the image. There is a strong relationship between the size of an object and the size of Gaussian kernel. Usually when a Laplacian operator is applied to an image, there is a possibility that it exaggerates the effect of noise and will not provide any directional information. Consequently Gaussian pre-smoothing is conducted to reduce the effect of noise. The arbitrary z-crossings approach was therefore developed to enhance the performance of edge detection based on a second derivative. It makes the contouring dynamic by providing the facility to expand or shrink the contour easily (Woolford, Hankamer et al. 2007).

1.5.3.10 Additional algorithms

Fast marching methods are those in which a contour is initialized from a pre-chosen seed point and allowed to grow until a certain boundary condition is reached (Bajaj, Yu et al. 2003, Baker, Yu et al. 2006). Traditionally these were designed for single object segmentation. For the segmentation of multiple objects, a seed point must be chosen for each object but this can be problematic.

Another edge detector is the **3D recursive filter** (Monga, Deriche et al. 1991) which approximates the gradient of a given image by computing the impulse response recursively. It then applies an

edge-closing algorithm in a pseudo 3D fashion; i.e. it works in XY, YZ, and XZ planes separately, however this limits the ability to accurately detect the 3D structure of an object. For more objective annotation and contouring of microtubules and ER-Golgi membrane-like cellular structures in electron tomogram, template matching (Lebbink, Geerts et al. 2007) has proven to be very useful.

Similarly, more modern filters (Frangakis, Bohm et al. 2002, Zhu, Carragher et al. 2003, Adiga, Malladi et al. 2004, Mallick, Zhu et al. 2004) exhibit variable edge-detection performance at low SNR. However, for the most part, these algorithms have only been implemented in 2D and thus have limited utility for analysing 3D image volumes. Therefore performance on 3D data is largely untested to date. A true 3D filter, capable of using data from adjacent slices, offers the advantage that it provides additional information from either side of the 'focal' slice, thereby enabling enhanced noise suppression along with the detection of contiguous and legitimate structural details throughout the 3D image stack.

The Canny edge detector and bilateral edge filter outperformed all other techniques and while the two have the same level of performance, the bilateral edge filter (one adjustable parameter) outperforms the Canny edge detector (three adjustable parameters) when taking into account the number of adjustable parameters (Pantelic, Ericksson et al. 2007).

Specific approaches were especially developed for 3D segmentation of ET targeting membranes. For example Martinez-Scanchez et al (Martinez-Sanchez, Garcia et al. 2011) used local differential structures to segment membranes in tomograms. This method first isolates membrane information according to spatial scale by defining a membrane-like points, which closely represent membrane model in tomogram in local-scale. Those points are then further analysed in a global scale to determine which of them actually constitute the membrane. The final points were then integrated into an output model to the yield final segmentation. In 2013, (Martinez-Sanchez, Garcia et al. 2013) proposed a new computational framework based on the detection and characterisation of local maxima. This local information was used to cluster the ridges (i.e. local maxima), to yield the segmented tomogram. The same group (Martinez-Sanchez, Garcia et al. 2014) proposed another approach for membranes segmentation which used a 'tensor voting algorithm' (Tong, Tang et al. 2004), for the analysis of differential information through 3D space. As all nearby voxels having coherent structural information enhance each other, the structural information of voxels that belong to the same membrane may facilitate segmentation of high-resolution structural studies of tomograms of vitrified samples.

The 3D Canny edge detector is widely used as a benchmarking filter (three adjustable parameters) for 2D analysis and its implementation for 3D tomographic data is also available. The 3D recursive filter offers a simplified alternative to the 3D Canny filter, but it uses a 2D edge-tracking algorithm that is applied to each XY, YZ and XZ plane separately, which limits its ability to accurately detect

edges in a 3D environment. Therefore, we aimed to extend the bilateral edge filter from 2D to 3D to make it suitable for fast and accurate segmentation of 3D electron tomograms.

1.5.4 Novel and innovative 3D algorithms

Automated tomogram segmentation ideally requires algorithms that are *robust, accurate, intuitive, fast, generally applicable* (i.e. able to detect all objects in a cell) and *selective* (i.e. able to specifically identify a target object from this diverse set of objects). Currently none of the algorithms available are able to meet all of these criteria. Manual segmentation is for example robust but not fast (Section 2.2.5, Figure 2.5). Template matching algorithms can detect specific target molecules (e.g. ribosomes) but are time consuming and not generally applicable for the detection of all objects in the cell).

Semi automated approaches such as the 'drawing tool' and 'interpolator tool' plug-ins (Noske, Costin et al. 2008) have extended the manual tracing capabilities of the IMOD software package (Alber, Forster et al. 2008). On the one hand this tool yields 3D surface-rendered models in a manner similar to manual segmentation as the contours of every 5th slice are manually traced and then interpolated by the software. However this process remains time consuming, requires biological expertise and is subject to human subjectivity. It is however *generally applicable* and capable of being *selective*.

Gradient based edge detectors: In image processing, the Prewitt operator (Judith M. S. Prewitt 1970), Roberts cross edge detector (Roberts 1963) and Sobel edge detectors (Gonzalez 2002) are considered classical edge detectors based on image gradients. These classical filters are computationally inexpensive, but very sensitive to noise. High levels of noise may result in degradation of edges and loss of quality. Furthermore these edge detectors are not applicable for *selective* object identification, as they cannot classify objects based on their 3D characteristics.

Canny edge detector: The Canny edge detector is a multi-stage complex filter developed for edge detection (Canny 1986), which requires the manual optimisation of 3 variables making it difficult to automate and fails to detect true discontinuity (i.e. to check whether the discontinuity is due to noise or a true edge), and therefore can produce spurious and unstable boundaries of objects embedded in noise corrupted data (see Chapter 2). Furthermore it is not applicable for *selective* object identification.

The snake algorithm: The Snake algorithm is an active contouring algorithm. The user manually

draws a rough contour (snake) enclosing the structure of interest, and the snake interprets the contour semi-automatically by pushing the snake closer to the feature of interest (Kass, Witkin et al. 1988). So it is intuitive but manual, and requires longer computation times. It is not applicable for *selective* object identification

The watershed transform is a mathematical morphology-based segmentation technique, based on the concept of watershed and catchment basins (Roerdink and Meijster 2001). Over segmentation is a known drawback of standard watershed algorithms.

The work presented in this Chapters 2, 3 and 4 is therefore focuses on the sequential development of novel automated algorithms that are *robust, accurate, intuitive, fast, generally applicable* and *selective.* To achieve this a number of factors must be taken into consideration, which contribute to the novelty of the work presented.

Generally applicable: Currently most, if not all algorithms are either manual or semi-automated in their implementation. Automation requires the ability to use a generally applicable approach. As all detected objects have edges, edge detection, provides such a generally applicable approach. Consequently the 3D BLE, RAZA and RAZA^{PS} (Chapter 2, 3 and 4) are edge detectors, which are capable not only of detecting organelles (Figures 2.5, 3.3, 3.5) and membranes (Figures 3.8) but also macromolecular complexes (Figure 2.4, 3.6), and membrane proteins (Figures 3.4, 3.7).

Robustness, sensitivity and accuracy: Automated approaches must be robust. 3D algorithms have a theoretical advantage over 2D algorithms in this regard. This is illustrated for a 3x3x3 = 27-voxel kernel in which the focal pixel focused on an edge enables a 26 way connectivity (i.e. 27 voxels – the focal pixel). In contrast a 3x3 pixel 2D kernel only offers 8-pixel connectivity. In other words in a noise corrupted image, a 3D kernel can extract an edge based on 3D rather than 2D connectivity, thereby increasing accuracy and sensitivity (Figure 2.1, 2.2, 2.3, 3.4 and 3.8) beyond that of a corresponding 2D algorithm and therefore also increases robustness.

Speed: Manual segmentation of approximately 1% of a mammalian cell to the level of membranes and large macromolecular assemblies took approximately 9 months (Marsh, Mastronarde et al. 2001), highlighting the importance of increasing speed of segmentation. Using automated segmentation approaches developed in Chapter 2 (Section 2.2.5/Figure 2.5), a 4000 x increase in segmentation speed was achieved at a similar level of accuracy (Ali, Landsberg et al. 2012). The speed of processing is dependent not only on the efficiency of the algorithm itself, but on any

required pre-processing, the number of algorithm parameters that must be manually optimised prior to the automated run and post processing steps. While the 3D recursive filter can be automated, it requires pre-processing (file conversion) and performed poorly in noise levels typical for cryo-EM. In contrast the Canny edge detector requires adjustment of three parameters (x: the standard deviation of the Gaussian, y: the high hysteresis threshold and z: the low hysteresis threshold) making the process much more labour-intensive and less *intuitive*. The 3D BLE (Chapter 2) did not require pre-processing, achieved improved *sensitivity* and *accuracy* and is fully automated, not requiring manual parameter optimisation.

Specificity: Specificity is the ability to specifically select a target object (e.g. a ribosome from a multitude of segmented objects detected in a cellular tomogram). To date specific selection of target objects has generally been achieved using correlation based template matching (Bohm, Frangakis et al. 2000), feature based approach, convolution based template matching but in 3D this process is slow and impractical for automation as a huge number of templates would be required to detect all cellular objects and because many objects are unique and so not well suited for detection by template matching. Such methods are computationally intensive. Chapter 3 and 4 present the RAZA and RAZA^{PS} filter, which yields complete 3D contours for each identified object. Each contour is a mathematically defined object allowing its length, width, height, volume and surface area parameters to be measured. This is a significant advance as the automated measurement of these objects specific 'fingerprints' enables selective target identification, based upon them (Figures 4.3, 4.5, 4.7 and 4.8). Furthermore the implementation is novel in that object identification can be conducted despite the random orientation of the target objects in the cell.

Intuitive: The 3D BLE algorithm offers a powerful parameter-free filtration capability by eliminating the need for manual σ^2 optimisation, which is well suited for high-throughput automated segmentation. To date there is to my knowledge no method available that facilitates parameter-free segmentation of electron tomograms. RAZA offers the ability to adjust 1) sigma (σ, the width of the Gaussian that controls the degree of smoothing) and 2) the Z-crossings value (a threshold that defines which edges are detected) but as is shown in Figures 3.3, 3.4, 3.5, 3.6, 3.7 and 3.8, their adjustment is intuitive and automatable.

The 3D methods developed in this thesis and particularly RAZA and RAZA^{PS} therefore yield novel contour information that will likely prove advantageous for subsequent down-stream processing steps such, molecule extraction for sub-volume averaging and the docking of higher resolution structures determined by SPA, NMR, X-ray and electron crystallography into cellular tomograms.

At least without further processing available 2D methods do not enable automated docking as they only yield contours and not specific 3D object fingerprints to identify target contours for docking (see Section 5.1/ Figure 5.1).

1.6 Docking of 3D structures *in situ* and initiating a multi scale 3D atlas

Since the dawn of structural biology, biologists have long since used their knowledge of the structure and organization of individual molecules to better understand their function. Molecular mapping of whole cell structure offers the potential to provide additional insight into the dynamics of structures and the crowded macromolecular organisation in a cellular context. Extensive research, which can potentially guide the assembly of a comprehensive 3D library/atlas of photosynthetic embedded structures, is already available. For example, the active participation of thylakoid membrane complexes in photosynthesis has been shown (Drop, Webber-Birungi et al. 2014, Gold, Ieva et al. 2014), but the organization of specific complexes is still only loosely defined. A multi-scale structural analysis is needed to achieve a true pseudo-atomic 3D reconstruction of the thylakoid membrane, incorporating tomographic, SPA and atomic resolution crystallographic data.

Fortunately, atomic models are already available for many proteins (PDB) . These atomic coordinates can be used to generate electron density maps (Forster, Han et al. 2010). Jochen Bohm and his colleagues were among the first to demonstrate detection of macromolecules (thermosomes, 20S proteasome and *E. coli* GroEL) within real tomographic data using cross correlation-based template matching (Bohm, Frangakis et al. 2000). Two years later, Achilleas Frangakis published a fully automated template-matching algorithm, which was able to detect and identify crowded 20S proteasomes and thermosomes, in a tomogram of phantom cells (Frangakis, Bohm et al. 2002).

Fitting of atomic models within intermediate resolution density maps can aid interpretation of the structure of very large assemblies, particularly for multi-component structures. The usefulness of docking component structures into a cryo-EM map has been demonstrated through the development of a detailed pseudo-atomic model of the mammalian 80S ribosome at 8.7 Å resolution (Chandramouli, Topf et al. 2008). Niels Volkmann proposed docking approaches based on density-correlation measures between electron density of atomic models and electron density resolved by EM (Volkmann and Hanein 1999, Volkmann and Hanein 2003). Such correlation-based approaches are relatively slow but highly accurate. To extend docking accuracy, they demonstrated a confidence interval-based method for calculating correlation coefficients from molecular docking experiments (Volkmann 2009).

A common framework capable of integrating structural and computational biological data that can bridge the resolution gap between membrane proteins, large multi-subunit macromolecular structures, and the structural organization of membranes would represent a timely development. Computational techniques such as sophisticated noise reduction, segmentation and pattern recognition techniques provide the potential to visualize macromolecules in a close to native state. The performance of such approaches is dependent on the quality of raw data and resolution. Sub-tomogram analysis includes extraction, alignment, classification and averaging of tomographic sub-volumes. To extract all the particles from a tomogram is very complex and time consuming. Many automated and semi-automated particle-picking techniques (Frangakis, Bohm et al. 2002, Roseman 2003, Volkmann 2004, Nickell, Forster et al. 2005, Woolford, Hankamer et al. 2007, Forster, Pruggnaller et al. 2008, Scheres, Melero et al. 2009) have been developed and used in different biological applications (Ortiz, Forster et al. 2006). Using a combination of computational techniques, ribosomes and some unknown structures were detected in a cellular context within cryo-electron tomograms of frozen hydrated *D. discoideum* cells (Frangakis and Forster 2004). More recently, the combination of ET with 3D image classification and averaging has begun to provide unique opportunities to determine the structure and conformation of membrane proteins in intact cells (Bartesaghi and Subramaniam 2009, Brandt, Etchells et al. 2009). Reconstruction of 3D maps of two microtubule-kinesin motor complexes demonstrated that post tomographic averaging of identical sub-volumes extracted from tomograms can be used to produce 3D maps, identification conformations which were not possible using helical or other averaging based techniques (Cope, Gilbert et al. 2010). Sub tomogram averaging on cryo electron tomograms of PSII and ATPases has revealed their supramolecular arrangement in thylakoid membranes of spinach and pea (Daum, Nicastro et al. 2010). Methodological advances in sub-tomogram analysis are thus anticipated to be an increasingly valuable computational tool allowing structural biologists to capture structurally heterogeneous macromolecular complexes *in situ* (Hrabe and Förster 2011).

1.7 Summary

Photosynthesis is of fundamental biological importance as almost all life on Earth depends on it, either directly (e.g. plants) or indirectly as its products (e.g. food crops, biofuels and chemical feedstocks). As the global population rises from 7 to 9 billion people this dependence on the process of photosynthesis is expected to increase and this is already being seen as food prices rise with demand. Microalgae are increasingly considered an important option for the extension of future food, fuel and chemical feedstock production, as they can be produced on non-arable land in many cases using saline water and provide a number of other attractive qualities such as high turnover,

short culturing cycle, contribute to carbon capture and lower water demands than conventional bio-fuels (Posten and Schaub 2009, Wijffels and Barbosa 2010). By understanding the structure-function relationship of photosynthetic proteins of microalgae, we can optimize the efficiency of such systems.

The first step of all photosynthetic processes is light-capture. The so-called light-harvesting proteins capture incident light and the energy is then transferred as excitation energy to the reaction centres of Photosystems I and II, which drive the photosynthetic electron transport chain, the production of NADPH and ATP for CO_2 fixation and the downstream storage of chemical energy. As light capture is the first step of all crop, biofuel and by product production, the optimization of its efficiency is central for the development of economic systems. However, the photosynthetic machinery has a remarkably complex and dynamic architecture, which enables adaptation to constantly changing environmental conditions. Over the past decade, great progress has been made in resolving the atomic structures of individual photosynthetic proteins and complexes. To understand the complex interplay of these protein complexes *in vivo*, we now require pseudo-atomic resolution 3D atlases of the cells into which these complexes can be docked. By analogy, just as a manual is required to tune a car engine, a molecular resolution 3D atlas of the photosynthetic machinery would assist the tuning of light capture processes for high-efficiency microalgal biotechnology applications. Computational methods such as noise reduction, edge detection and segmentation algorithms are central to the development of such 3D atlases.

The focus of my PhD project is to develop and evaluate computational tools enabling the production of 3D maps of the photosynthetic machinery at molecular resolution within thylakoid membranes. Biological samples are very sensitive to beam/radiation damage. Therefore, low dose imaging conditions and the effects of the missing wedge (Figure 1.23) result in low SNR, distortion and anisotropic resolution. Sample preparation also introduces low and non-uniform contrast and complicates the process of extraction and interpretation of the structural features from highly complex and inherently noisy electron tomograms. CCD image capture also introduces additional noise and undesirable artifacts. The combination of noise, low SNR and low contrast pose significant hurdles, which limit accurate 3D edge detection.

1.8 Thesis objectives

Based on the above I therefore aim to develop new computational techniques capable of simultaneous noise suppression and edge detection of macromolecules and membrane protein complexes in tomograms. The specific aims/objectives of this research project are listed below:

Aim 1: “To evaluate (review) available imaging technologies. To identify how available methods can best be used to perform noise reduction and image segmentation in 3D to help define the structure of macromolecular objects”. *This objective is addressed in Chapter 1.*

Aim 2: “To develop, test and evaluate a 3D noise reduction and edge detection filter that is capable of resolving the contours of individual macromolecular assemblies in the cell”. *This objective is addressed in Chapter 2 & 3.*

Aim 3: “To compare manual segmentation with semi-automated segmentation facilitated by edge detection methods”. *This objective is addressed in Chapter 2 & 3.*

Aim 4: “Detect membrane proteins within tomographic data using edge detection methods”. *This objective is addressed in Chapter 3, Figure 3.4 and 3.7*

Aim 5: “To use edge detection methods to develop 3D object selection tools for the quantification of macromolecular assemblies and subcellular compartments by defining *structural fingerprints* using geometric properties of segmented objects”. *This objective is addressed in Chapter 4.*
However, the ultimate goal is to reveal functionally important properties from analyses of the structures of protein complexes involved in the photosynthesis process, based on information gathered from the combination of ET, SPA and electron crystallographic data. This integrative analysis provides the means to resolve the remarkable supramolecular architecture of the dynamic, solar-powered photosynthetic machinery within thylakoid membranes.

1.9 Thesis structure

The remainder of this thesis is organized as follows. **Chapter 2** outlines a novel ‘3D Bilateral edges detector’ developed and used to facilitate rapid automatic edge detection of macromolecular complexes. **Chapter 3** illustrates another novel ‘RAZA edge detector’ to improve the final quality and overall accuracy for segmentation of whole cell volumes. **Chapter 4** presents a particle detection technique ‘$RAZA^{PS}$’ for analysis of 3D tomograms and the results of the detailed comparative analyses of the structural details (height, width, length, surface area and volume) of detected sub-volumes. Finally, **Chapter 5** discusses both the technical and biological outcomes from this project in the context of future directions, and concludes the thesis with a summary of major findings.

CHAPTER 2

3D Bilateral Edge Filter (3D BLE)

3D image reconstruction of large cellular volumes by electron tomography (ET) at high (≤5 nm) resolution can now routinely resolve organellar and compartmental membrane structures, protein coats, cyto-skeletal filaments and macromolecules. However, current image analysis methods for identifying *in situ* macromolecular structures within the crowded 3D ultrastructural landscape of a cell remain labour intensive, time-consuming and prone to user-bias and/or error. This paper demonstrates the development and application of a parameter-free, 3D implementation of the bilateral edge-detection (BLE) algorithm for the rapid and accurate segmentation of cellular tomograms. The performance of the 3D BLE filter has been tested on a range of synthetic and real biological data sets and validated against current leading filters – the pseudo 3D recursive and Canny filters. The performance of the 3D BLE filter was found to be comparable to or better than that of both the 3D recursive and Canny filters while offering the significant advantage that it requires no parameter input or optimisation. Edge widths as little as 2 pixels are reproducibly detected with signal intensity and grey scale values as low as 0.72 % above the mean of the background noise. The 3D BLE thus provides an efficient method for the automated segmentation of complex cellular structures across multiple scales for further downstream processing, such as cellular annotation and sub-tomogram averaging and provides a valuable tool for the accurate and high- throughput identification and annotation of 3D structural complexity at the subcellular level, as well as for mapping the spatial and temporal rearrangement of macromolecular assemblies *in situ* within cellular tomograms.

2.1 Introduction

Electron tomography (ET) is an important tool for studying structural cell biology *in situ* by bridging the resolution gap between light microscopy and methods for protein structure determination at atomic resolution, such as x-ray and electron crystallography as well as nuclear magnetic resonance (NMR) spectroscopy. Recent advances in ET at the level of sample preparation, improved detector sensitivity/capture efficiency and imaging resolution, along with automated

computational techniques for 3D image reconstruction, processing and analysis now enable macromolecular assemblies to be resolved at up to 15-30 Å, in the best case examples (Hrabe and Förster 2011). Meanwhile, the 3D reconstruction of extremely large cytoplasmic volumes at ~3-6 nm resolution and even entire mammalian cells at ~10 nm resolution by cellular ET now affords unprecedented new insights regarding the structure-function relationships that exist among subcellular compartments/organelles, the plasma membrane, cytoskeletal filaments, large macromolecular assemblies as well as membrane proteins (Baumeister 2002, Marsh, Volkmann et al. 2004, Daum, Nicastro et al. 2010, Volkmann 2010). 3D cellular reconstructions of this nature thus provide a precise spatial framework for developing annotated, pseudo-atomic resolution 3D atlases of cells through docking high resolution structures of macromolecular assemblies.

A critical step in the advancement of molecular resolution ET is the ability to accurately segment molecular structures *in situ* within cellular tomograms. Classical edge-detection algorithms such as the Sobel (Gonzalez 2002), Prewitt (Gonzalez 2002), Laplacian of Gaussian (Gonzalez 2002) and Canny edge detectors (John 1986) are increasingly being incorporated into semi-automated and automated methods for segmenting 3D image volumes. However, all of these are best suited to images with relatively high signal-to-noise ratios (SNR) and thus have limited use for the accurate/automated analysis of cellular tomograms, which have an inherently low SNR. By comparison, more modern filters (Volkmann 2002, Zhu, Carragher et al. 2003, Adiga, Malladi et al. 2004, Mallick, Zhu et al. 2004, Yu and Bajaj 2004, Adiga, Baxter et al. 2005, Pantelic, Ericksson et al. 2007, Woolford, Ericksson et al. 2007, Woolford, Hankamer et al. 2007) exhibit improved edge-detection performance at low SNR. However, for the most part these algorithms have only been implemented in 2D and thus have limited utility for analysing 3D image volumes. A true 3D filter, capable of using data from adjacent slices, offers the advantage that additional information from either side of the 'focal' slice can be considered, thereby enabling enhanced noise suppression along with the detection of contiguous and legitimate structural details throughout the 3D image stack.

The Canny edge detector (John 1986, Sonka, Hlavac et al. 1999) is widely considered to be a "gold standard" filter (Jin 2006) for 2D analysis. More recently it has been implemented in 3D (*http://www.imagescience.org/meijering/software/featurej/edges.html*). This implementation is a multi-stage, complex filter, which in principle involves four fundamental steps. In the first step it convolves the target volume with a Gaussian filter to smooth the image and suppress the noise. The second step calculates gradients of the image using a Sobel edge detector, the rationale of applying which is to identify voxels with sufficiently large weighting magnitudes that identify them as an edge. In step three, non-maximum peak suppression is performed to track the edge points along the high magnitude regions and to eliminate the remaining voxels. This is followed by step four, which

through hysteresis thresholding converts the output volume into a binary format to ensure that noise voxels are not included as part of a true edge.

Optimization of the filter's performance requires the simultaneous fine-tuning of three parameters: the standard deviation of the Gaussian as well as the high and low hysteresis thresholds. The need to simultaneously optimize multiple parameters makes the use of the 3D Canny labor-intensive and impractical for application to high throughput, automated or semi-automated analysis. An additional drawback is that it fails to detect true discontinuities (i.e. it is unable to discriminate contrast discontinuities that are due to noise or a true edge).

The 3D recursive filter (Monga 1991) offers a simplified alternative to the 3D Canny filter. It approximates the gradient of an image by computing the impulse response recursively and finally applies a pseudo 3D edge-closing algorithm; that is, it uses a 2D edge-tracking algorithm (Deriche, Cocquerez et al. 1988) that is applied to each XY, XZ and YZ plane separately. In practice, the tracking algorithm, which is designed to complete discontinuous contours within a 2D plane, limits its ability to accurately detect the 3D structure of an object.

The underlying principles of the bilateral edge filter (Pantelic, Ericksson et al. 2007) offer an attractive alternative to the Canny and 3D recursive edge-detection filters. The BLE is a nonlinear, photometrically-weighted, discontinuity-based anisotropic filtering technique that has been shown to be suited to images containing predominantly low- and mid-frequency information (Pantelic, Ericksson et al. 2007). More specifically, it suppresses noise by attenuating undesired frequencies and enhances edge-detection by selectively extracting specific features. However it still requires user modification of the manual parameter (σ_2) and has not been implemented in 3D.

In this chapter, we present a full 3D implementation of the bilateral edge filter (3D BLE), which importantly also eliminates the need for manual σ_2 optimization. This fully automated 3D BLE is a simple and fast filter, specifically designed for electron tomography data which typically has low signal to noise ratios, but can also be used to analyse a wide range of other 3D data. Our implementation includes Gaussian filtration followed by iterative median filtration as a pre-filtering step (Gonzalez 2002, van der Heide, Xu et al. 2007). Iterative median filter maintains edges and converges to an optimal solution beyond which it does not keep smoothing. This significantly improves edge-detection by simplifying the voxel intensity distribution and suppressing noise disturbances without corrupting the edge information. The output 3D binary image data can then be used together with automated segmentation or for edge-detection-based 3D particle picking.

2.2 Results

2.2.1 Adaptation of the 2D bilateral edge filter to analysis of 3D image volumes

In this paper, the original 2D variant of the BLE filter (Pantelic, Ericksson et al. 2007) has been extended to operate in three dimensions. Additionally, parameter adjustment has been fully automated. Similar to the 2D BLE filter, the 3D BLE filter first calculates the photometric score for each individual voxel (focal voxel) in the context of the 'processing window' of the image volume being analysed. A score of 0 represents a perfect edge while 1 represents noise, for each individual voxel. The rationale of sequentially calculating photometric scores for each focal voxel, as the window moves across the image, is to build a photometric score map, from which edges can be traced. The photometric function $\Phi(x,y,z,m,n,o)$, corresponding to an adjacent neighboring voxel (m,n,o) from the focal voxel (x,y,z) is defined in Equation 2.1.

$$\boldsymbol{\Phi(x,y,z,m,n,o) = e^{-\frac{|I(x+m,y+n,z+o)-I(x,y,z)|^2}{2\sigma_2^2}}} \qquad (\mathbf{2.1})$$

$I(x,y,z)$ is the original volume and (x,y,z) indicates the coordinates of the focal voxel, while (m,n,o) indicates the adjacent coordinates of a neighbouring voxel to the focal voxel (which can be expressed as; m={ x-1, x, x+1 }, n={ y-1,y,y+1 } and o={ z-1,z,z+1 }). σ_2 is a photometric parameter which defines the minimum difference in intensity that is to be regarded as an edge. In the 2D BLE, σ_2 is the only parameter requiring manual adjustment. While manual adjustment ensures the highest quality result as it allows the fine tuning of σ_2, in this 3D implementation the user can choose to adjust this parameter automatically to achieve high throughput, by replacing it with the actual standard deviation (σ) of the volume (Equation 2. 2):

$$\sigma = \sqrt{1/N \sum_{i=1}^{N} (X_i - \bar{X})^2} \qquad (2.2)$$

where N is the total number of voxels, X_i is the current voxel intensity and $\bar{X}$ is the mean value of all the intensities present in the volume. Automated optimisation of this photometric parameter is highly desirable to facilitate segmentation of large and complex cellular tomograms in a high throughput manner.

Voxels having intensities above the background level may correspond to signal or noise and the further such a high intensity voxel is from a focal voxel, located on a given edge, the less likely it is

to be part of that edge. Consequently, to distinguish between edge and noise voxels, each photometric score of a neighbouring voxel (m,n,o) is next spatially weighted according to a 3D Gaussian distribution centred at the focal voxel (i.e. the further a voxel of a given intensity is from the focal voxel, the lower its photometrically-weighted score). The Gaussian weightings $\gamma(m,n,o)$ are given in Equation 2.4.

$$\gamma(m,n,o) = e^{-\frac{m^2+n^2+o^2}{2\sigma_1^2}} \qquad (2.3)$$

c

Here σ_1 is used to define the significant weights and is fixed to 2 voxels, similar to the 2D case (Pantelic, Ericksson et al. 2007). It is also used to determine the size of the neighborhood for the calculation of the normalized photometric score given in Equation 2.4.

To identify substantial discontinuity (i.e. an edge), the photometric score is next normalized by averaging across a pre-defined number of voxels (Equation 2.5). The normalized photometric score $\overline{\Phi}(x,y,z)$ for the given focal voxel is calculated as a Gaussian weighted average of the individual photometric scores over a radius of 2 voxels.

$$\overline{\boldsymbol{\Phi}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \frac{\sum_{m=-2}^{2}\sum_{n=-2}^{2}\sum_{o=-2}^{2}\boldsymbol{\gamma}(\boldsymbol{m},\boldsymbol{n},\boldsymbol{o})\boldsymbol{\Phi}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{m},\mathbf{n},\mathbf{o})}{\sum_{m=-2}^{2}\sum_{n=-2}^{2}\sum_{o=-2}^{2}\boldsymbol{\gamma}(\boldsymbol{m},\boldsymbol{n},\boldsymbol{o})} \qquad \mathbf{(2.4)}$$

The normalized photometric scores indicate the significance of discontinuity of a given edge. Scores close to 1 are considered weak photometric responses and can correspond to fluctuations in background noise, single, high intensity, spurious voxels, or the focal voxel otherwise not being centered on an edge. Strong scores (close to 0) represent voxels that are likely part of an edge. After analysing the possible numerical representations of connectivity of an edge a threshold of 0.85 was chosen. This setting was repeatedly found to provide the largest observed gain in edge-detection performance. Average photometric scores below this threshold are considered to indicate edge voxels and scores greater than 0.85 indicate spurious voxels, non-edge voxels, or weak fluctuations in background intensity due to anisotropy. The thresholded photometric score $\overline{\Phi}'(x,y,z)$ is thus given by Equation 2.6:

$$\overline{\Phi'}(x,y,z) = \begin{cases} \overline{\Phi}: \overline{\Phi}(x,y,z) < 0.85, \\ 1: \overline{\Phi}(x,y,z) \geq 0.85 \end{cases} \qquad (2.6)$$

Consequently $\overline{\Phi}'(x,y,z)$ remains unaltered if its value is below 0.85, but is adjusted to 1 if its value is greater than or equal to 0.85. The lowest scores in the final photometric map generated by Equation 2.5 represent local minima or troughs along which an edge is traced based on the use of a 3×3×3 voxel volume. A focal voxel in a 3×3×3 volume (i.e. 27 voxels) therefore has the possibility of forming 26 different connections with its immediate neighbors. If the focal voxel is part of a continuous trough (edge) or at the end of a trough, its average photometric score will be within the thirteen smallest scores of a (3×3×3) voxel neighborhood. This is because twelve represents the number of voxels required to describe a continuous edge within this 3D volume of 27 voxels (4 voxels within each plane of a 3×3×3 volume = 12 voxels). Increasing this value tends to increase edge connectivity further in noise free data, but in the presence of high levels of noise (e.g. cryo-EM data) this tends to result in decreased noise suppression. Theoretically, a continuous edge of 5 voxels within a 3×3 voxel plane (= 15 voxels in a 3×3×3 volume) is possible. However thresholding at a value of 15 also allows edges with branching (or noise contamination) to be detected. Thirteen in our experience is a sensible compromise (see Supplementary Material of Pantelic *et al.* (Pantelic, Ericksson et al. 2007) for further explanation).

2.2.2 Application to synthetic data

Using a combination of real and synthetic datasets, the performance of the 3D BLE filter was evaluated and compared to two benchmark 3D edge-detection algorithms: the 3D Canny and 3D recursive filters. To obtain a fair comparison, all test volumes in the initial phase were subjected to the same pre-filtering step and the recommended or default settings of each filter were used. No attempts at parameter adjustment were made. As a first test, the noise suppression and edge-detection abilities of the 3D BLE filter were evaluated using synthetic "truth" reference volumes (hollow cylindrical, spherical, triangular and rectangular) contaminated with different combinations of Gaussian (G) and impulse (I) noise (Figure 2.1). Results were evaluated based on three criteria: response to true edge directionality, the minimum detectable object edge width (in pixels) and capability to detect true edges and distinguish them from noise in images corrupted with high levels of noise (Figure 2.1, Table 2.1). Such an analysis enables the quantification of the minimum possible signal to noise ratio (SNR) required for the filter to detect edges.

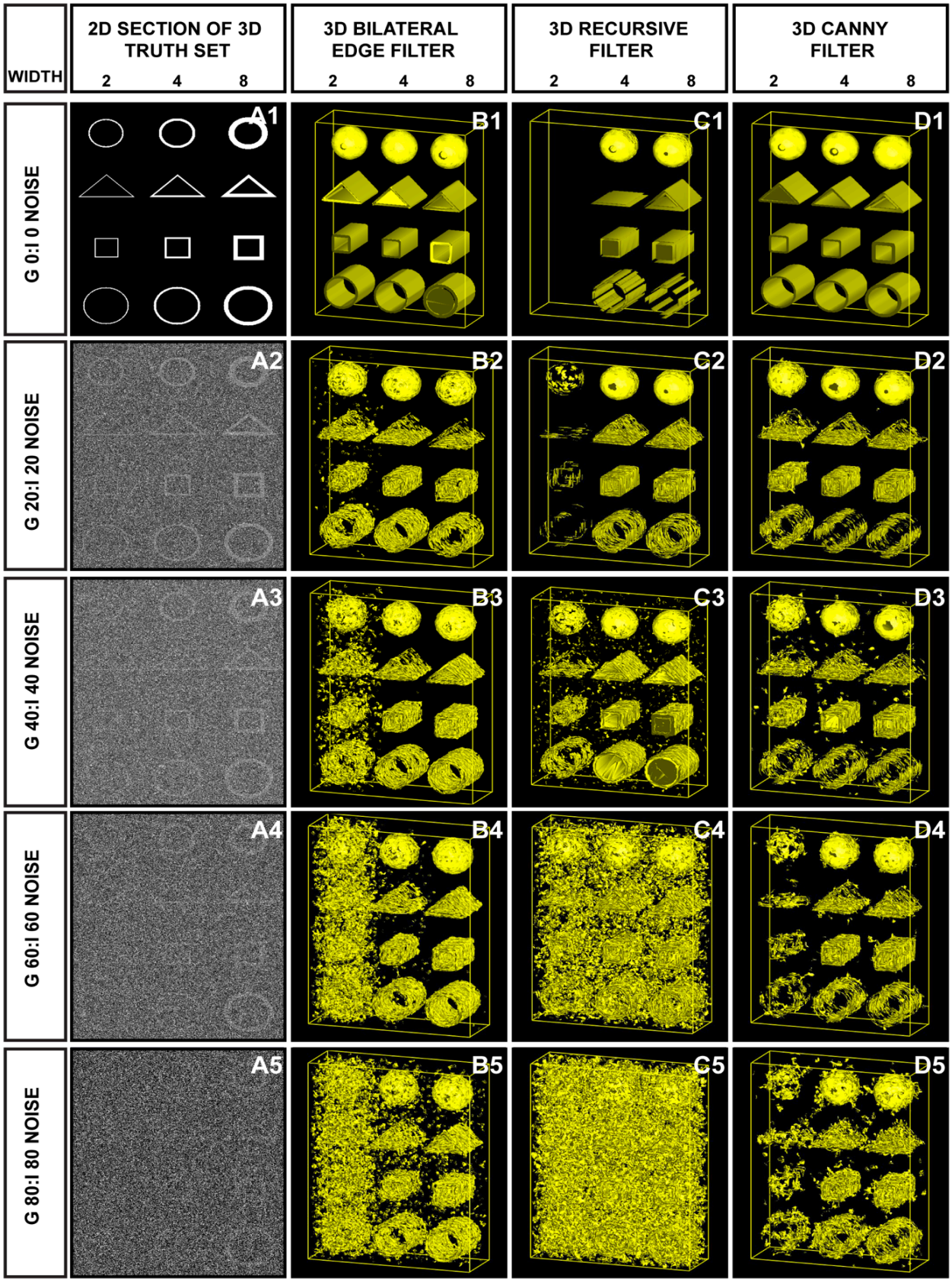


Figure 2.1 Application of 3D BLE to synthetic phantoms with Gaussian and impulse noise.

Performance of the 3D BLE, 3D recursive and 3D Canny filters was assessed using a volume of 3D synthetic phantoms contaminated with increasing levels of Gaussian and impulse noise. **(A1-A5)** 2D sections taken from synthetic volumes contaminated with increasing levels of Gaussian and impulse noise. **(B1-B5)** 3D

Surface rendering of results (B1-B5) obtained from the 3D BLE filter. **(C1-C5)** Surface rendering of the 3D recursive-filtered synthetic dataset. **(D1-D5)** Surface rendering of the 3D Canny-filtered synthetic dataset.

Noise was introduced incrementally from an initial value of 5% G/5% I (expressed as a percentage of the signal intensity) up to G 80%/I 80%. The width of the reference objects was varied from 1 to 16 pixels. Figure 2.1 shows a subsample of the tests performed that highlight the performance limitations of the 3D BLE (Figure 2.1B) in comparison to the 3D recursive (Figure 2.1C) and Canny filters (Figure 2.1D). Post-processed 3D surface views of the test datasets clearly highlight performance differences between the three filters. At low noise levels (G 0/I 0 to G 20/I 20) the three filters performed similarly and effectively detected the edges of all reference volumes. However as the noise levels increased further (G 40/I 40 to G 80/I 80) differences in performance emerged. At G 60/I 60 or greater, the 3D BLE is unable to effectively discriminate the structure of the object from noise at an edge width of 2 pixels, but is able to recognise all test objects with an edge width of 4 pixels or greater even at G 80/I 80. Corruption with G 80/I 80 noise provides a stringent test in which it is almost impossible to distinguish the objects from noise by eye (Figure 2.1 A5) but the 3D BLE is still capable of significantly amplifying object information above the level of the noise for edge widths of 4 and 8 pixels. By comparison, the capabilities of the 3D recursive filter appear to be limited at G 60/I 60 and beyond, regardless of edge width. Furthermore at the lower noise levels of G 40/I 40 and even G 20/I 20 the 3D recursive filter, while effectively dampening background noise, poorly resolved edges 2 pixels in width in comparison with the 3D BLE and the Canny filters. Consequently we conclude that the performance of the 3D BLE filter was better than that of the 3D recursive filter and particularly so at high noise levels.

A comparison of the performance of the 3D BLE filter with the Canny filter at G 60/I 60 and above indicates that the 3D BLE achieves a better edge connectivity at 4-8 pixel widths (see Figure 2.1 B4 vs. D4; B5 vs. D5). In the 2 pixel test the 3D BLE showed signs of noise contamination (Figure 2.1 B4 & B5) while the Canny filter showed poor connectivity (Figure 2.1 D4 & D5). At G 80/I 80, the Canny and 3D recursive edge detectors both treat high frequency noise voxels as edges and thus fail to highlight true structural information (see Figure 2.1 C5-D5). In contrast the 3D BLE filter, while retaining some high frequency noise voxels, appears to better preserve edge detail (Figure 2.1 B5). Collectively these tests suggest that the 3D BLE filter has a similar level of performance to the Canny filter. An important advance over the Canny, however, is that the 3D BLE filter can achieve a similar level of performance for many test images of this type in an automated fashion, while the Canny requires the optimization of 3 parameters for each image. Parameter-free filtration is essential for automated segmentation.

Noise Gaussian/Impulse	3D BLE (%)	3D recursive (%)	3D Canny (%)
0/0	**29.66** (1.35)	**30.03** (1.39)	**26.91** (1.11)
20/20	**46.24** (3.20)	**118.46** (21.58)	**38.52** (2.28)
40/40	**52.52** (4.20)	**120.74** (22.42)	**42.05** (2.72)
60/60	**63.36** (6.17)	**123.00** (23.27)	**43.73** (2.94)
80/80	**67.86** (7.08)	**124.96** (24.01)	**45.62** (3.20)

Table 2.1 Statistical evaluation of filter performance

The performance is evaluated by using synthetic volumes contaminated with different levels of Gaussian and impulse noise shown in Figure 2.1. RMSE scores between the input volumes and the three filter outputs are shown in bold. Smaller scores represent higher levels of correlation with the input volume. Values in brackets are the percentage voxel variation between input volumes and the three filtered outputs.

Table 2.1 lists the respective root mean square error (RMSE) values calculated between the filtered volumes and the truth set. Analogous to Pantelic *et al.* (Pantelic, Ericksson et al. 2007), the truth image was constructed by applying the Canny edge-detector to the noise/CTF/envelope-free variant of Figure 2.1 A1. We then compared the truth image with the results generated by all filters (Figure 2.1 B1-5 to D1-5). The 3D BLE filter outperformed the recursive filter, yielding lower RMSE values in each case (see Table. 1). However at all five tests the Canny yielded the lowest RMSE scores. The apparently improved performance of the Canny filter under these conditions is not surprising due to the fact that the truth image is constructed based on the output of the Canny edge detector. The percentage voxel variation between the truth image and the filtered image is also calculated (see Table. 2.1) and the improved performance of the Canny filter under these conditions can be explained by the more rapid fall off in edge-detection by the 3D BLE for thin objects (edge width of 2 pixels) compared with that of the Canny, where performance deteriorated almost equally, regardless of edge width. In summary, these tests indicated that the 3D BLE filter performed significantly better than the 3D recursive filter and similarly to the Canny filter under the test conditions analysed.

The preceding tests indicated that the 3D BLE filter is able to resolve contours of truth reference images from synthetic impulse and Gaussian noise at levels up to 80% of the signal intensity. In the next stage of testing the truth reference set was contaminated with noise more closely simulating "real" experimental conditions encountered in electron micrographs, in order to provide a more realistic evaluation of the performance of the 3D BLE filter for EM data.

The results of these analyses provided a clear indication of the limitations of each of the three filters tested. Unsurprisingly, for case III (perfectly overlayed noise and signal), none of the three filters were able to distinguish signal from noise (nor did any of the three filters show any detection

artefacts). In all other cases, the 3D BLE filter performed at least as well or better than both the 3D recursive and Canny algorithms.

At the less stringent settings (C1-E1; C5-E5) all three filters detected 4 and 8 pixel wide edges. The Canny filter appeared to be more affected by background noise (Figure 2.2 C1 vs E1; C5 vs E5) but also at the cost of artifactual noise contamination of the edge. It did however detect 2 pixel edges better. In contrast the 3D BLE filter yielded clearly defined and relatively noise-free contours but resolved the 2 pixel edges less well. Connectivity can be increased manually in the 3D BLE filter by raising the threshold setting from 0.85 (see Equation 2.5) but at a cost of increased background noise. This suggests that the Canny and 3D BLE filters have a similar level of performance.

At the most stringent intensity test settings (B2-E2; B4-E4) the 3D BLE and Canny filters both performed significantly better than the 3D recursive filter. The Canny and 3D BLE filters both effectively detected 4-8 pixel wide edges. Again the Canny filter yielded a level of performance better than 3D BLE but at the cost of increased artifactual noise contamination of the edge. In terms of the detection of 2 pixel edges, the 3D BLE and Canny filters were both close to their limits of detection. Generally the 3D BLE appeared to be better at suppressing noise while the Canny appeared better at detecting 2 pixel edges, but the differences are minimal.

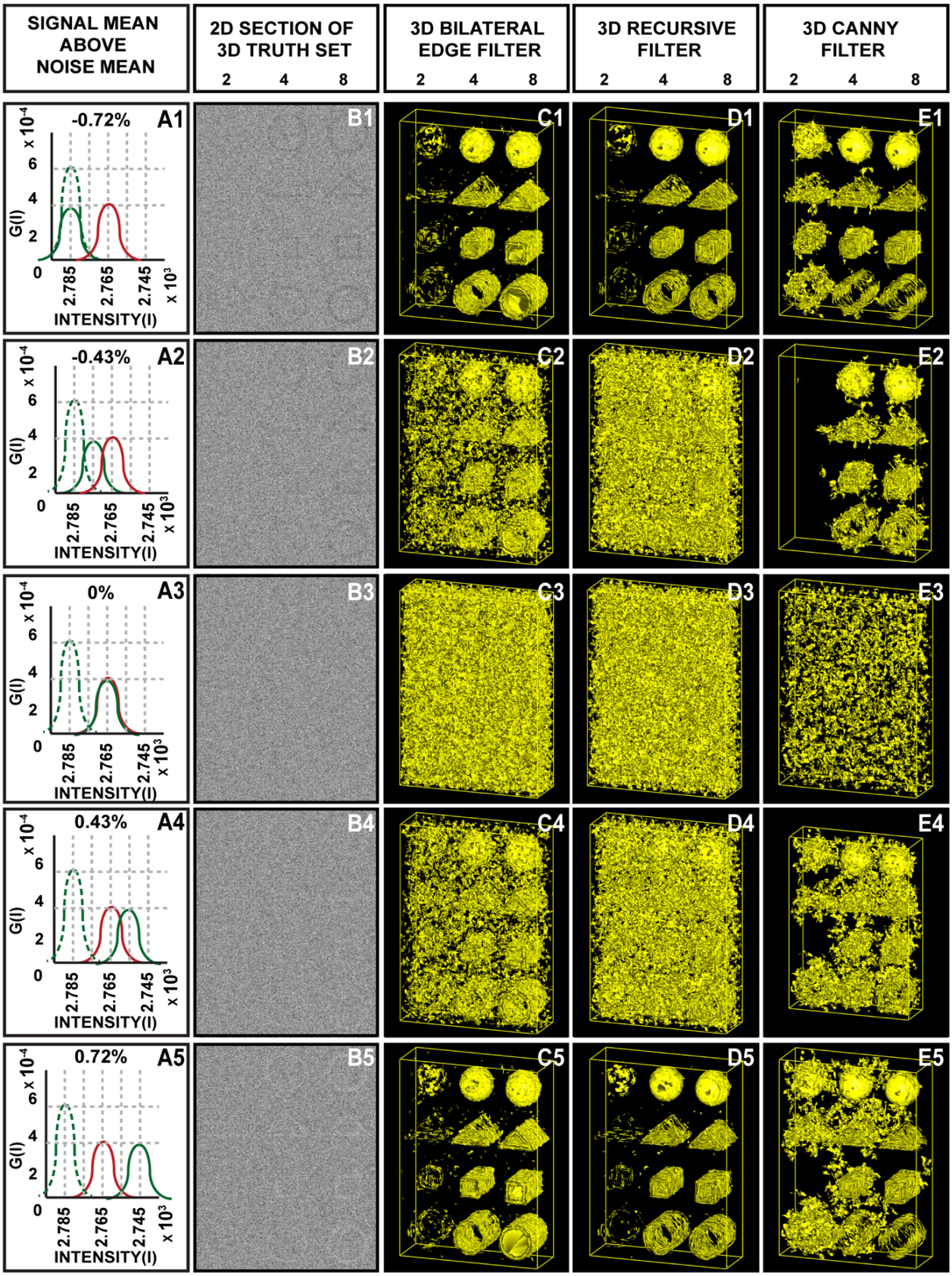


Figure 2.2 Application of 3D BLE to synthetic phantoms with simulated cytosolic noise.

Performance of the 3D BLE, 3D recursive and 3D Canny filters was assessed using the same volume of 3D synthetic phantoms shown in Figure 2.1, but contaminated with different levels of simulated experimental noise. (A1–A5) A graphical representation of the SNR present in the five representative cases shown. Coloring in A1–A5 is as follows: green dotted line shows the contrast and intensity of the original signal; red line shows the contrast and intensity of the noise; green solid line shows the scaling and shifting of signal

profile towards noise profile. Overall, the graph shows the probability density function (G(I)) of the normal distribution (B1–B5) 2D sections taken from synthetic volumes contaminated with experimental noise. (C1–C5) 3D surface rendering of results obtained following application of the 3D BLE filter to the synthetic dataset. (D1–D5) Surface rendering of 3D recursive- filtered test dataset. (E1–E5) Surface rendering of 3D Canny-filtered dataset.

Performance of the three filters was quantified by calculating RMSE values (Table 2.2) between the control/truth volume constructed by applying the Canny edge-detector to the noise/CTF/envelope-free variant of Figure 2.1-A1 and the filtered output volumes (Figure 2.2 C-E). The 3D Canny outperformed the 3D BLE when the signal mean was offset ±0.43% from the noise, but failed to recover the edge information in all cases especially at an object width of 2 pixels and signal offset of -0.43%. The 3D BLE outperformed the 3D recursive at ±0.43%, but performed roughly equal to the 3D recursive at ±0.72% with minor differences in RMSE scores.

Noise Signal mean above noise mean	3D BLE (%)	3D recursive (%)	3D Canny (%)
-0.72	48.8 (3.66)	44.47 (3.04)	48.61 (3.63)
-0.43	65.13 (6.52)	73.08 (8.21)	48.82 (3.66)
0	99.48 (15.22)	80.42 (9.95)	58.67 (5.29)
0.43	65.51 (6.60)	71.22 (7.80)	58.89 (5.33)
0.72	49.28 (3.74)	44.64 (3.07)	53.82 (4.46)

Table 2.2 Statistical evaluation of filter performance

The statistical evaluation is performed by using synthetic volumes contaminated with different levels of simulated experimental noise shown in Figure 2.2. RMSE scores between the input volumes and the three filter outputs are shown in bold. Smaller scores represent higher levels of correlation with the input volume. Values in brackets are the percentage voxel variation between input volumes and the three filtered outputs.

2.2.3 Application to biological test data

Having evaluated the performance of the 3D BLE filter using a synthetic truth reference set comprised of simple geometric shapes, we next used a simulated cryo-tomogram populated with 100 uniquely oriented copies of the GroEL chaperonin complex to evaluate the performance of the filter in a more biologically relevant context.

Figure 2.3 demonstrates the performance of the three filters on a representative area of this simulated cryo-tomogram (Ludtke, Chen et al. 2004) containing nine GroEL molecules in different orientations. The first column (Figure 2.3 A1-A5) shows the relative ratio and offset of the mean signal contrast compared to the background noise (as in Figure 2.2 A1-A5). Upon close inspection, the 3D BLE (Figure 2.3 C1-5) and 3D Canny (Figure 2.3 E1-5) filters show better edge connectivity than the 3D recursive (Figure 2.3 D1-5). In addition, the 3D recursive filter starts to become less effective at removing noise at ±0.43% noise over signal (Figure 2.3 D2, D4). The 3D BLE filter, with fully automated parameter optimisation, performed nearly as well as the 3D Canny. Differences in the performance of the 3D BLE and Canny filters are attributed to connectivity of additional noise in the Canny images and lower detection of 2 pixel edges in the BLE. Overall therefore the BLE and Canny again seemed similar in performance, but the fully automated 3D BLE filter provided a considerable processing advantage.

2.2.4 Application to experimental data

We compared the ability of the 3D BLE, 3D recursive and 3D Canny filters to extract molecular edge contours from an 800 x 800 x 100 voxel region of a dual tilt tomographic reconstruction recorded from a resin-embedded, sectioned and post-stained *C. reinhardtii* cell (Figure 2.4). The dark densities are putative macromolecular assemblies, having an approximate diameter of 25 nm – roughly equivalent to the size of a ribosome. In this test, the 3D BLE filter (Figure 2.4B) clearly outperformed both the 3D recursive (Figure 2.4C) and 3D Canny edge detector (Figure 2.4D). Noise suppression in the 3D BLE-filtered image (Figure 2.4B) was considerably enhanced and contours around the putative macromolecular particles are thus more accurate and less corrupted by spurious noise densities. The likely reason for the 3D BLE filter outperforming the Canny and 3D recursive filters in this test is that the SNR

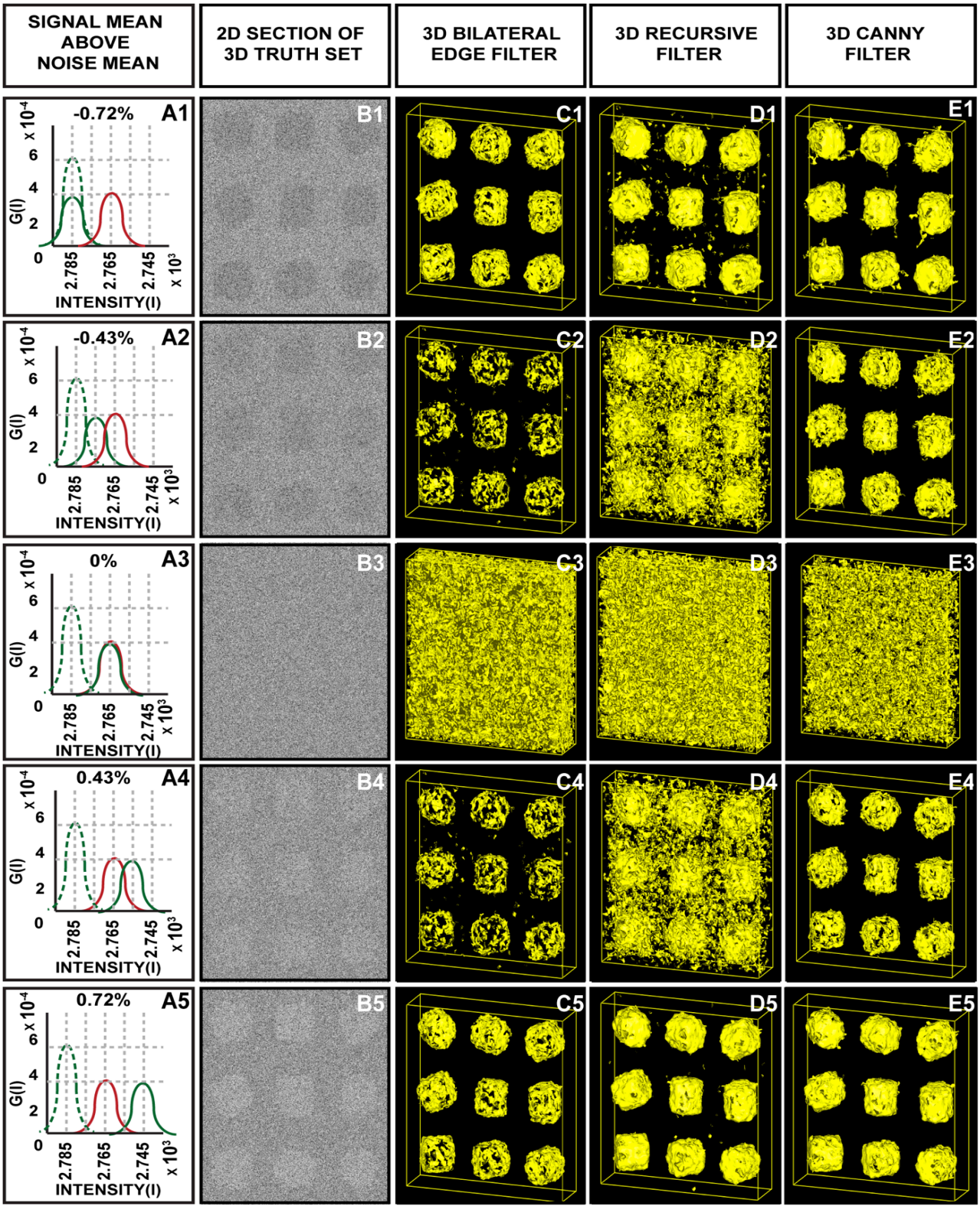


Figure 2.3 Detection of molecular volumes using 3D BLE.

The ability of the 3D BLE, 3D recursive and 3D Canny filters to resolve molecular contours was assessed using a test volume populated with 3D GroEL molecules. A representative region of the test volume showing 9 molecules is shown. (A1-A5) SNR illustrated as for Figure 2.2. (B1-B5) 2D sections taken from synthetic volumes contaminated with experimental noise. (C1-C5) Surface rendering of results following application of the 3D BLE filter applied to the test volume contaminated with experimental noise. (D1-D5) Surface rendering of 3D recursive-filtered test volume. (E1-E5) Surface rendering of 3D Canny-filtered test volume.

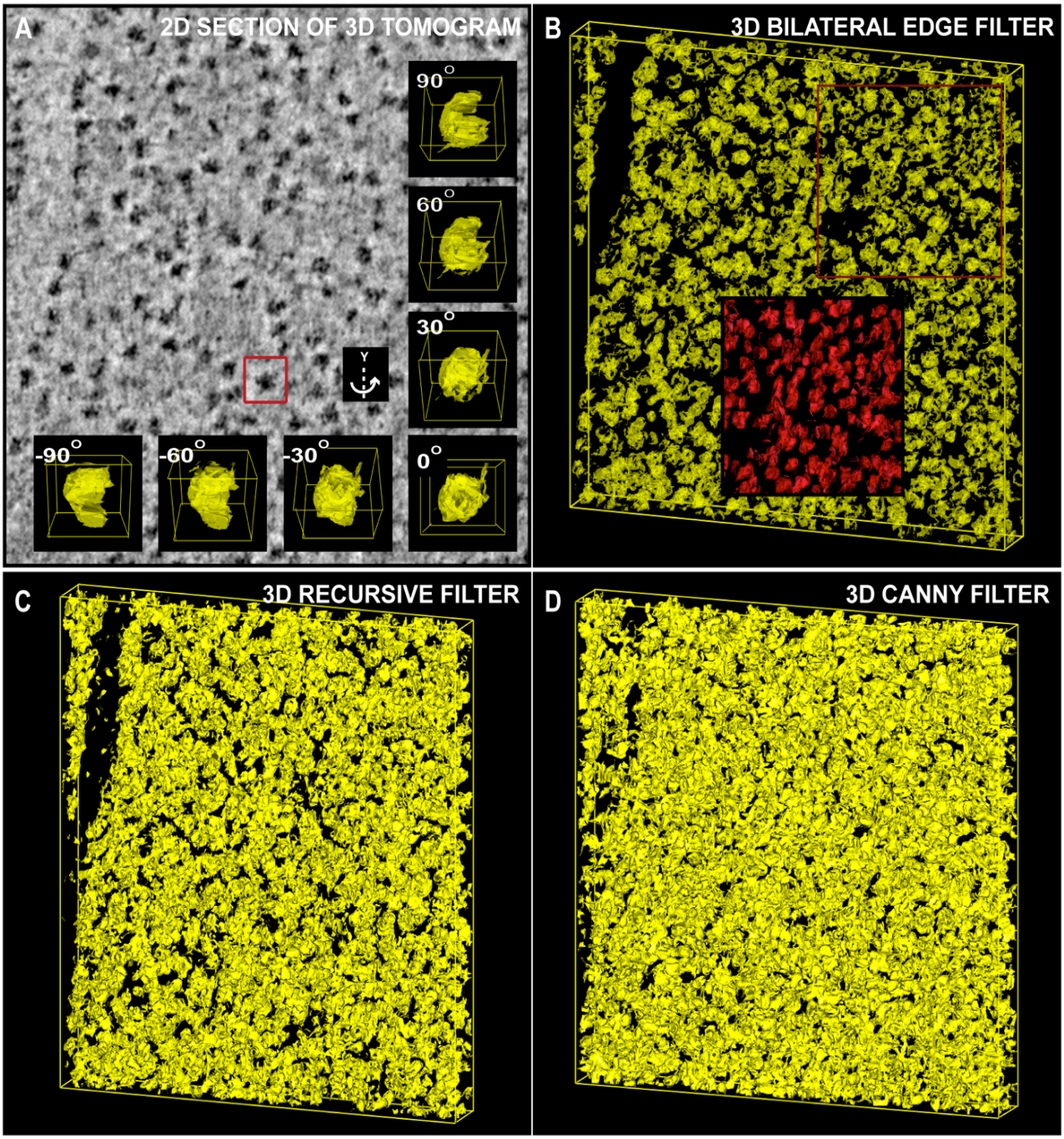


Figure 2.4 Extraction of molecular contours from an electron tomogram subvolume.

Application of the 3D BLE, 3D recursive and 3D Canny filters to a subvolume of an experimentally-recorded tomograms of a resin-embedded *Chlamydomonas reinhardtii* cell. (A) Unprocessed, central 2D cross-section of the subvolume extracted from the 3D tomogram showing a region of the chloroplast heavily populated with putative macromolecular assemblies (dark objects). The inset in (A) highlights a randomly chosen single particle, represented at as an isosurface rendering and shown at a selected number of orientations around the y-axis. (B) 3D surface rendering of results obtained from application of the 3D BLE filter. (C) Surface rendering of the 3D recursive-filtered subvolume. (D) Surface rendering of the 3D Canny-filtered subvolume.

of the input data (Figure 2.4A) is significantly higher than that of any of the other test images (see Figure 2.1 & 2.3) and was consistent with the enhanced performance of the 3D BLE filter under high SNR conditions in all of the previous tests (see Figure 2.1 & 2.2; Tables 1 & 2). Consequently,

we conclude that the 3D BLE behaved better by detecting more true positive edges and suppressed more noise for this tomographic data set, and that it again achieved this using a fully automated algorithm.

The ultimate aim of running an edge-detection algorithm is to obtain high quality and continuous 3D contours. Noise can however result in discontinuities. To enhance the performance of the 3D BLE filter further in this regard, the edges that it detected were completed using a Bspline interpolation (Figure 2.4B-inset). The use of the Bspline clearly improved connectivity. This can be seen at the molecular scale in Figure 2.4A (insets) where one of the segmented particle volumes is shown in a range of orientations. This experiment demonstrated that the accuracy of the molecular contours obtained was high and sufficient for the detection of individual macromolecular assemblies within experimentally-recorded electron tomograms. This potentially paves the way for automated detection and extraction of molecular volumes for downstream 3D alignment, classification and single particle averaging.

2.2.5 Automation of the development of a molecular resolution cellular 3D atlas

In the preceding sections we established that the automated 3D BLE filter can be applied to recover molecular-level detail from noise-corrupted and real experimental volumes. In particular we demonstrated the application of the 3D BLE filter to the accurate and automated segmentation of macromolecular structures *in situ*. While the examples used in the preceding sections are based on real tomographic data, the detection of individual particles was not complicated by the presence of neighboring structures that might confound its performance. As a final test, we therefore applied the 3D BLE algorithm to a larger subvolume packed with potentially confounding macromolecular and organellar structures. For this test, we extracted a subvolume from a tomogram encompassing a large cytoplasmic volume imaged from a murine pancreatic cell that had been analysed in detail by manual segmentation in a previous study (Marsh, Mastronarde et al. 2001). This set was chosen because it contained tightly packed molecular and organellar contours and because the performance of the 3D BLE could be compared to that achieved by manual segmentation performed as part of the original study.

The results obtained following application of the 3D BLE to the pancreatic cell tomogram are shown in Figure 2.5. Figure 2.5A shows a representative 2D section of the 3D tomogram. Figure 2.5B shows the manually segmented structures reported by Marsh *et al.* (Marsh, Mastronarde et al. 2001) for a representative subvolume (demarcated by the red box in Figure 2.5A) of the full 3D tomogram. Figure 2.5C shows a 3D surface view of Figure 2.5B. Figure 2.5D shows the output obtained following processing of the complete volume with the 3D BLE filter. Figure 2.5E shows the corresponding densities detected using the 3D BLE filter. A comparison of Figure 2.5B and E

indicates that 3D BLE filter was able to detect all of the structural elements identified by manual segmentation in a fully automated manner. The tomographic slice in Figure 2.5G (which is extracted from the dataset used in Figure 2.5A) shows the membrane organization of a mitochondrion in the region. Figure 2.5H shows the surface view of the outer membrane as well as the inner cristae (pink). The detection of inner cristae clearly highlights the capability of the 3D BLE, where manual segmentation is labor intensive and requires some biological expertise.

These results indicate that at the organellar level, the 3D BLE is capable of extracting structural data from cellular electron tomograms in an automated manner. The speed with which this was achieved is another important property of this filter. The 3.1 x 3.2 x 1.2 μm^3 volume reported by Marsh *et al.* (Marsh, Mastronarde et al. 2001) required around 9-12 months to segment manually using IMOD (Mastronarde 1997). The 3D BLE detected all of these structures (see Fig 6B, full volume) in 1 h 53 min, which is approximately 4000x faster. It should also be noted that the 3D BLE additionally detected a much more extensive set of particle contours (Figure 2.5D), but only those that corresponded to the set of previously reported manual contours were shown, in order to facilitate a direct comparison. The current implementation of the 3D BLE required that the detected contours corresponding to those reported by Marsh *et al.* (Marsh, Mastronarde et al. 2001) be marked up (i.e. colored) manually, a process which remains time consuming. In the future however the use of algorithms such as those reported by Woolford *et al.* (Woolford, Hankamer et al. 2007) could be used to detect and therefore classify particles/organelles based on size and shape to establish rules for a first pass of automated contour color application prior to manual curation.

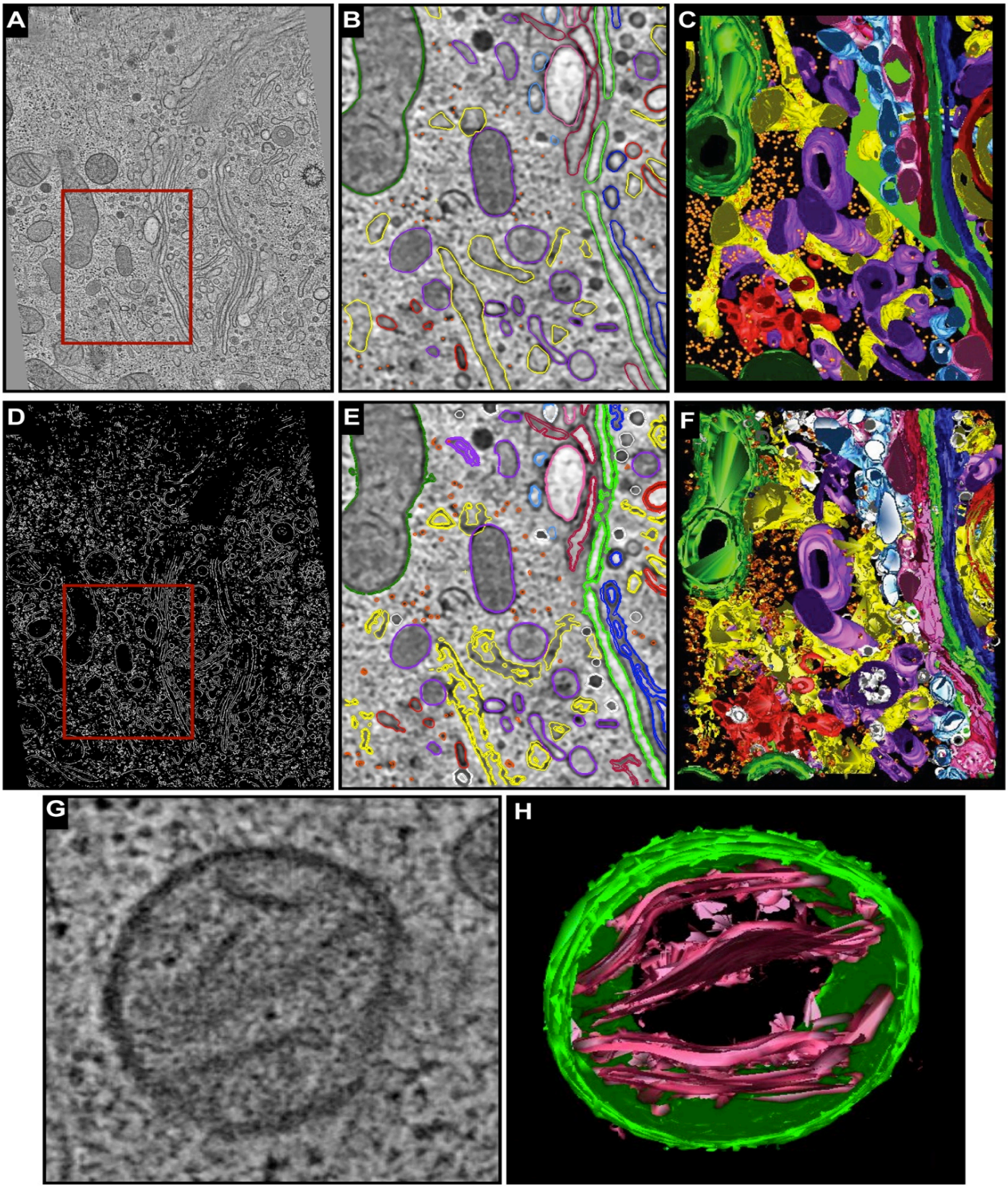


Figure 2.5 Segmenting Golgi region of an insulin-secreting pancreatic beta cell line.

(A) A tomographic slice (slice 33) extracted from the reconstructed volume reported in (Marsh, Mastronarde et al. 2001). The region demarcated by a red box is shown in (B). (B) Objects were segmented by manually drawing colored lines (contours) using IMOD. (C) Surface-rendered 3D model of the Golgi region analysed in B by manual segmentation. (D) 3D BLE-filtered tomogram. (E) Contours detected automatically by the 3D BLE were then manually colored for comparison to the manually segmented volume shown in (B). (F) Surface-rendered 3D model generated by automatic segmentation of the same region shown in B. Coloring in (C-D) and (E-F) is as follows: the seven cisternae that comprise the Golgi in the region - C1, light blue; C2, pink; C3, cherry red; C4, green; C5, dark blue; C6, gold; C7, bright red. ER, yellow; membrane-bound ribosomes, blue; free ribosomes, orange; mitochondria, bright green; dense core vesicles, bright blue; clathrin-negative vesicles, white; clathrin-positive compartments and vesicles, bright red; clathrin-negative

compartments and vesicles, purple; mitochondria, dark green. (G) A tomographic slice revealing the outer and inner membrane architecture of a mitochondrion in the Golgi region. (H) Surface rendering shows that automated 3D segmentation facilitated by the application of 3D BLE detects the mitochondrial membranes.

2.2.6 Evaluation of computational requirements

In addition to noise suppression and true edge accuracy, an important consideration in comparing the 3D BLE filter to current gold standard filters is the consumption of computational resources. Table 2.3 summarises the processing times and memory usage for the 3D BLE, 3D recursive and Canny filters. When considering the processing time required by the filter alone, the 3D BLE at first appeared to perform worst of all three filters, taking 63 s of CPU time to process the 385 x 512 x 128 voxel test volume shown in Figure 2.1, compared with 11 s for the 3D recursive filter and 21 s for the Canny. But the 3D BLE filter compared much more favourably, when taking into account the requisite pre- and post-processing steps. The 3D BLE filter reported here is implemented with a small, fixed filter window (5 x 5 local neighborhood - experimentally determined to minimize window size) with pre-processing and post-processing routines incorporated directly into the filter algorithm. In comparison, the 3D recursive filter required several pre-processing steps including the conversion of the input file to a raw image format as well as post-processing which included thresholding and conversion back to the original (in this example, MRC) image format. Taking these additional steps into account, the processing time of the 3D BLE (169 s) was less than that of the 3D recursive filter (186 s) and as has already been shown, yielded considerably improved results. The 3D BLE also required slightly less memory than the recursive filter when the resource requirements for file conversion and thresholding were taken into account (240.8 Mb vs ~272 Mb).

	CPU time Total[a] (filter) sec	Memory Mb	No. of adjustable parameters
3D BLE	169 (63)	240.8	0
3D recursive	186 (11)	221.6 (272)	0
3D Canny	21[b]	485.0	3

Table 2.3 Comparison of processing resources consumed by each of the three filters.

All filters are evaluated using a synthetic volume (385 x 512 x 128 voxels) contaminated with different levels of Gaussian and impulse noise.

[a] Total processing time includes pre-processing and post processing.

[b] Although the run time of the Canny filter is short, the process is not automated.

The Canny filter required no pre- or post-processing but unlike the 3D BLE and 3D recursive filters, which are fully automated, the Canny edge detector has three adjustable parameters (x: the

standard deviation of the Gaussian, y: the high hysteresis threshold and z: the low hysteresis threshold), which require optimisation. To evaluate all combinations of just two different values for each parameter would require 8 iterations of the filter. In practice we have found that in the best cases, a minimum of 10 parameter combinations had to be tested to yield a result comparable in quality to that of the 3D BLE. Correspondingly, the effective processing time of the 3D Canny is increased at least 10-fold from 21 s to ~210 s. This represents an increase of approximately 15-20% over and above the processing times of the 3D BLE and 3D recursive edge detectors ((210/187) x 100 = 15%) when pre- and post-processing were taken into account, while making the process much more labor intensive. Memory requirements (485 Mb) of the 3D Canny were also substantially higher – approximately double that of the 3D BLE. This highlights the value of the fully automated 3D BLE filter, in particular for high volume and/or high throughput image processing, where the one "adjustable" parameter (σ_2) is automatically optimised in our implementation.

2.3 Conclusion

We have described here a *bona fide* 3D implementation of the BLE filter that is able to accurately recover 3D contours describing the structure of individual macromolecular assemblies within real tomographic reconstructions of subcellular volumes. In these tests, the ability of the 3D BLE to accurately localize and detect edges in conjunction with noise suppression is demonstrated. The performance meets or surpasses that of computationally more expensive 3D edge detectors by providing a straightforward and automatable implementation that does not require manual parameter adjustment. It is especially well suited to 3D particle detection for subsequent volume extraction, 3D alignment and averaging and thus holds great promise for the rapid and accurate segmentation/identification of 3D macromolecular structures. The fact that the algorithm also yields contour information will likely prove advantageous for subsequent down-stream processing steps such as docking higher resolution structures determined from SPA, NMR, x-ray and electron crystallography *in situ* within lower resolution cellular tomograms.

2.4 Design and implementation

2.4.1 Implementation

The 3D bilateral edge filter was developed in C++ using the BSoft C++ library (Heymann and Belnap 2007, Heymann, Cardone et al. 2008). The code has been compiled and tested on Mac OS X operating systems (Snow Leopard). All testing and experiments were conducted on 3D volumes.

2.4.2 Test data/patterns

The accuracy and integrity of the implemented 3D BLE was initially tested using a truth set comprising a broad range of conceptual reference volumes representing different geometries (sphere, cylinder, triangular and rectangular prism) and edge widths (1, 2, 4, 8 or 16 pixels) designed to thoroughly test the response of the filter to curves, straight lines and directionality. The reference volumes were then sequentially corrupted with increasing combinations of impulse and Gaussian noise (5-80%), or with simulated experimental noise (see below). Filter performance was assessed in comparison to leading filters in the field including the 3D Canny (John 1986) and 3D recursive filters (Monga 1991).

Noise suppression and edge-detection capabilities of the 3D bilateral edge filter were also evaluated using an experimental test volume populated with one hundred uniquely oriented density maps generated from the 6 Å 3D reconstruction of GroEL (EMDB accession code 1081) (Ludtke, Chen et al. 2004). In order to comprehensively evaluate filter detection limits, the test volume was contaminated with differing amounts of simulated experimental noise adjusted to achieve mean signal intensity either greater, equal or less than the mean noise intensity. Contrast variation was normally distributed by adjusting the signal mean intensity (-8.6% to +8.6%) to match that of the noise mean intensity. The width of the distribution was parameterised by the standard deviation of signal (-47% to +47%) to match that of the noise standard deviation, where 0% was the mean contrast of the embedded test objects.

The simulated experimental noise was extracted from cytosolic regions of an algal cell tomogram in which organelles, filaments or other major subcellular structures were absent. The intensity and contrast profiles of the noise were defined and this information was then used to model the noise profile in Figure 2.2 and 2.3. The signal strength of the truth reference particles was normalised relative to the noise i.e. the mean intensity and contrast of the truth reference images was set to the same value as the mean values of the noise so that particles were initially undetectable (Figure 2.3 B3) and the contrast of the truth sets then adjusted in order to identify the maximum experimental noise tolerated by the three edge detectors.

Filter performance was evaluated in terms of true and false positive object detection rates and the calculated RMSE between the filtered volumes and the corresponding uncontaminated original volumes, as well as by comparison of processing times and memory requirements.

2.4.3 Electron tomography

The final suite of tests was performed on electron tomograms of either the chloroplast region of a *C. reinhardtii* cell or the Golgi region of an insulin-secreting pancreatic cell. For *C. Reinhardtii*, cells (strain stm3 (Mussgnug, Thomas-Hall et al. 2007)) were prepared for plastic embedding by

concentration, high pressure freezing then freeze substitution and fixation using 2% OsO_4 (osmium tetroxide) and 1% TA (tannic acid) according to Jimenez *et al.* (Jimenez, Vocking et al. 2009). 300 nm sections were cut and post stained using 2% aqueous uranyl acetate and Reynolds lead citrate. Sections were imaged at 23,000x nominal magnification using a Tecnai F30 FEG-TEM (FEI) operating at 300 kV, equipped with a 4Kx4K lens-coupled camera (Direct Electron). Tilt series data were collected over a range of ±60° at 1.5° increments along two orthogonal axes and recombined computationally to produce a dual-axis 3D reconstruction using the IMOD software package (Kremer, Mastronarde et al. 1996, Mastronarde 1997, Marsh, Mastronarde et al. 2001). Experiments detailing the imaging, reconstruction and manual segmentation of the Golgi region of an insulin-secreting pancreatic are the focus of a separate study reported previously by Marsh *et al.* (Marsh, Mastronarde et al. 2001).

2.4.4 Shortcomings of 3D bilateral edge filter

Despite the advantages of the 3D BLE algorithm, it had a number of limitations. One of the major drawbacks of the 3D BLE algorithm is that when applied to noise contaminated datasets such as cryo-tomograms, closed contours cannot be guaranteed. Consequently the final segmented objects, which are essentially voxel demarcated rather than continuous contours, cannot be presented as unique 3D mathematical objects. In order to close the contours, the resulting edges have to be connected manually. The reason that this is important is that when discontinuous contours are generated, the geometrical calculations of object specific parameters including height, width, length, surface area, and volumes cannot be conducted. This in turn has downstream implications for object detection, segmentation, identification and annotation. Another issue is that the 3D BLE algorithm generates a binary output (where 255 is an edge and 0 is not an edge). Consequently when multiple objects in a tomogram are identified they cannot be classified. This too prevents, identification and annotation, as well as object specific sub-volume extraction for sub-tomogram averaging. It also results in the fact that for colour rendition, each object colour must be manually corrected using additional tools.

CHAPTER 3

RAZA: Rapid, Automated 3D Z-Crossings Algorithm

The visualization of the 3D ultrastructure of cells at atomic resolution is a major goal of cellular and structural biology. So far, no single technique can achieve this and the most advanced processes available require the integration (docking) of multi-scale data sets collected at high resolutions via electron tomography (ET), single particle analysis (SPA), crystallography and NMR. ET enables whole cells, and their constituent organelles and macromolecular assemblies to be visualized *in situ* to a resolution of ~2-5nm, while SPA, crystallography and NMR provide atomic resolution detail of individual protein complexes. Despite these advances the accurate location, segmentation and annotation of the myriad of intricately arranged cellular components remains challenging. This is due to the low contrast and structural complexity of the data and that manual approaches are considered as the 'gold standard'.

In the previous chapter, I addressed my research goal "to develop, test and evaluate a 3D noise reduction and edge detection filter to resolve the contours of individual macromolecular assemblies in the cell". The preceding chapter, therefore demonstrates the development of the 3D BLE filter, however in highly noisy tomographic data this filter does not produce continuous edges. An object with continuous edges can be used to further accurately define its geometrical properties such as volume and surface area. In this chapter, I again address the same research goal by developing a novel edge detector RAZA, that produces continuous edges and more accurate than 3D BLE. Note that 3D BLE has an advantage of facilitating automated edge detection. Furthermore, similar to Chapter 2, I aim to compare manual segmentation with semi-automated segmentation facilitated by edge detection algorithms. The results shown in the preceding chapter are primarily based on resin-embedded negatively stained samples. However cryo-electron tomograms help in detecting macromolecular complexes in their functional cellular context. Therefore, I will also show how

RAZA can be used to segment membrane proteins in cryo-electron tomogram, which is another goal of this thesis.

Therefore Chapter 3 describes the development and implementation of RAZA (The Rapid, Automated 3D Z-crossings Algorithm) which enables rapid, automated and selective detection and mathematical segmentation of organelles, membrane proteins and macromolecular assemblies using an integrated 3D Laplacian of Gaussian (LoG) / arbitrary Z-crossing approach. Each segmented object is defined by its height, width, length, volume and surface area. These object-specific parameters enable selective object identification, contouring, and molecular extraction for downstream applications, such as sub-volume tomogram averaging. Furthermore accurate, automated contouring of individual macromolecular assemblies and membrane proteins provides the high-throughput approach required to facilitate docking of atomic resolution protein structures into tomograms, yielding pseudo-atomic resolution 3D cellular reconstructions in a timely manner.

3.1 Introduction

Resolving the 3D architecture of cells on an atomic scale is one of the most ambitious challenges of cellular and structural biology, and is accompanied by the promise of delivering unprecedented insight into the complex and dynamic interplay between organelles, subcellular structures, macromolecular assemblies and biomolecules. Currently no single structural technique can achieve this aim, but the use of a set of biophysical techniques spanning the atomic to micron range (e.g. optical microscopy, cellular and single particle electron tomography, biomolecular NMR, X-ray and electron crystallography) the potential to yield 'pseudo-atomic' resolution 3D cellular atlases based on nested, multiscale datasets (Alber, Forster et al. 2008). The techniques of X-ray and electron crystallography, NMR, single particle analysis and molecular docking are well established in terms of their ability to recover atomic resolution (10^{-10} m) detail for soluble proteins, membrane proteins and macromolecular assemblies (Alber, Forster et al. 2008, Hellmich and Gaudet 2014, Shi 2014). Similarly, advanced optical, super resolution microscopy technologies have enabled the recovery of data in the 10^{-3}-10^{-9} m range (Schermelleh, Heintzmann et al. 2010). However, the routine recovery of cellular data in the 2-5 nm resolution range, required to resolve macromolecular assemblies and large membrane proteins in their cellular context, can currently only be achieved using electron tomography. These techniques have in recent years undergone significant improvements in sample preparation and instrumentation (Denk and Horstmann 2004, Liao, Cao et al. 2013, Villa, Schaffer et al. 2013, Amunts, Brown et al. 2014, Kuhlbrandt 2014, Pantelic, Fu et al. 2014, Ramachandra, Bouwer et al. 2014), allowing structural biologists to resolve not only organelle and membrane structures, but cytosolic macromolecular assemblies, cellular motors and the extrinsic densities of large membrane proteins embedded in native membranes in unprecedented detail (Daum, Nicastro

et al. 2010, Landsberg, Jones et al. 2011, Liao, Cao et al. 2013, Amunts, Brown et al. 2014, Kuhlbrandt 2014, Pantelic, Fu et al. 2014). The advent of direct electron detectors (DED), single electron counting and super resolution capabilities for electron microscopy has significantly increased obtainable signal-to-noise Ratio (SNR) across the spectral range which are reflected in higher detective quantum efficiency (DQE) (McMullan, Faruqi et al. 2014). Despite these improvements the accurate and automated detection of contours that define discrete structures of cellular and molecular features remains a computational challenge, as does the high-throughput selective extraction of molecular tomograms for sub-volume averaging and molecular annotation and docking.

The ability to accurately detect edges within electron tomographic data is paramount for molecular segmentation. Edges are regions in which there is a significant change in voxel intensity and also offer the benefit of enabling segmentation of objects regardless of their size (e.g. organelles and macromolecular assemblies), shape, relative orientation or contour greyscale value. Consequently edge detection algorithms can be used to geometrically define structural features across multiple scales, and facilitate semi- or fully-automated segmentation of information-dense 3D tomographic volumes.

Edge detection algorithms commonly reported in the literature can be classified as belonging to one of three categories. Classical edge detectors such as the Sobel (Rafael C. Gonzalez 2002), Prewitt (Judith M. S. Prewitt 1970) and Roberts edge detectors (Roberts 1963) rely on the definition *a priori* of the *maximum rate of change* of intensity (1st order derivative) that defines an edge. These algorithms can detect the location of an edge and its directionality, but due to a lack of a smoothing component are very sensitive to noise, and Often yield non-specific edge connectivities. Consequently, classical edged detectors of this type are mainly suited for applications dealing with data that has considerably higher SNR than typical ET data.

A second category of edge detectors includes the Canny edge detector (Canny 1986), bilateral filter (Tomasi and Manduchi 1998) and bilateral edge filter (Pantelic, Ericksson et al. 2007). These Gaussian-based detectors generally outperform 1st order derivative based filters when applied to data with SNR more typical of electron tomography data, but do not deal well with edges exhibiting strong directional change due to the loss of detail inherent in Gaussian smoothing due to sensitivity to noise. For example, in the case of densely populated tomograms, Gaussian based-filters do not completely differentiate between real structure and artifacts. However through careful optimisation of the parametric settings, in 2D, the Canny edge detector and bilateral edge filter were able to outperform other techniques (Pantelic, Ericksson et al. 2007). Furthermore while these two filters exhibited comparable levels of performance, the bilateral edge filter has the advantage of requiring

the adjustment of only one parameter rather than three, making it more intuitive to use and effective at extracting structural data from low SNR electron tomograms (Pantelic, Ericksson et al. 2007).

A third class of detectors include a Laplace operator (2nd order derivative i.e. *the rate of change of the gradient of intensity*) such as Laplacian of Gaussian (LoG) (Rafael C. Gonzalez 2002), Laplacian of Gaussian and arbitrary Z-crossings (Woolford, Ericksson et al. 2007, Woolford, Hankamer et al. 2007) and the Marr-Hildreth filters (Marr and Hildreth 1980). The ability to measure the rate of change of the gradient (i.e., Laplacian) of intensity, increases their ability to select between real and spurious edges as noise is less noticeable, thereby improving detection results.

Some algorithms (Monga, Deriche et al. 1991) act as sequential 2D filters and lack a full 3D implementation, which is less than ideal for high throughput automated 3D image analysis. To address this we recently reported a 3D implementation of the bilateral edge detector (Ali, Landsberg et al. 2012) which offered the added advantage of eliminating the requirement for interactive parameter adjustment for the detection of large macromolecular assemblies within tomograms. While the filter performed well, this 3D bilateral edge filter produced disconnected edges rather than single mathematical objects when the data had a low SNR and suffered from noise corruption. This discontinuity of the contours negatively effects the final reconstruction of a 3D surface. Note that the estimation of geometric properties (e.g., orthogonal axis lengths, surface area and volume) is either complicated or not possible when contours are incomplete and this in turn limits the ability to define the centers of objects and the extraction of large datasets of target objects (e.g. ribosomes) for subsequent sub-volume averaging e.g. Dynamo (Castano-Diez, Kudryashev et al. 2012), PyTom (Hrabe, Chen et al. 2012), PEET (Nicastro, Schwartz et al. 2006, Cope, Gilbert et al. 2010) .

Here we describe the Rapid Automated Z-crossings Algorithm (RAZA), a new algorithm for the semi-automated processing of low SNR data typical of ET (resin embedded and cryo), which overcomes these problems. RAZA combines the advantages of the Gaussian denoising operation with an adaptive edge detection capability for the full or selective automated edge detection of organelles, membranes, macromolecular assemblies and even large membrane proteins. In particular, RAZA utilises the fact that all resolved subcellular structures are defined by an edge and that edges are themselves all characterized by a significant change in voxel intensity. This change in voxel intensity can be used to detect edges, which represent the boundaries of such objects and upon edge detection, discrete volumes can be segmented (e.g. organelles, membranes, macromolecular assemblies). Consequently RAZA yields a 3D edge map, which contains all the information related to the detected edges and their positions. It is an autonomous non-contextual approach, which does not require prior knowledge of the objects and edges. Another major advantage of this approach is that it produces continuous defined edges by providing user-defined

flexible threshold (see Method section). The continuous contour defining a discrete object allows the software to estimate specific geometrical features including orthogonal axis lengths, surface area and volume. These in turn provide mathematical constraints and significant advantages for advanced structural cell biology. For example, it allows the centres of the 3D object to be identified enabling the extraction of its volume for sub-volume averaging. Furthermore, complete contouring assists with accurate molecular docking into tomograms and so with the production of 3D pseudo-atomic resolution cell reconstructions from multi-scale data sets.

3.2 Methods

3.2.1 Electron Tomographic Datasets

The accuracy and robustness of RAZA was tested on four ET datasets, representing both vitrified and resin-embedded samples.

3.2.1.1 High-pressure frozen, freeze-substituted, plastic-embedded samples:

The first dataset was a whole cell tomogram reconstruction of a high-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell as described previously (Noske, Costin et al. 2008). Briefly, Islets of Langerhans were isolated and purified and 3D reconstructions of beta cells were obtained using a Tecnai F30 electron microscope (FEI Company) operating at 300 keV under low dose conditions. This dataset resolved key subcellular organelles of the insulin-secretory pathways, including the Golgi apparatus, mitochondria, insulin secretory granules and multi-granular bodies. The dataset represents a rigorous test for the RAZA filter as it contains a broad range of grayscale values ranging from close to white (e.g. Golgi), light gray (e.g. mitochondria) and dark gray/black organelles (e.g. insulin secretory granules).

A second tomogram used for testing was a tomogram of a high pressure frozen, freeze substituted, resin embedded Golgi region of an insulin-secreting pancreatic beta cell. Sections were imaged at either 15,700x using a JEM-4000FX (400 kV) intermediate voltage EM (JEOL), or at 14,690x nominal magnification using a JEM-1000 high voltage EM (at 750 kV (JEOL)). Tilt series data were collected over a range of 120° at 1.5° increments along two orthogonal axes and recorded on film. These tilt series were then recombined computationally to produce a dual-axis 3D reconstruction using the IMOD software package (Kremer, Mastronarde et al. 1996). Three serial 400 nm sections were aligned with each other, to produce a total reconstructed volume measuring 3.1 x 4.7 x 1.2 μm^3. More details on imaging and reconstruction to produce this dataset are described in a previous study (Marsh, Mastronarde et al.).

A third test was performed on a tomographic reconstruction of the chloroplast region of a high pressure frozen, freeze substituted plastic-embedded *C. reinhardtii* cell (Ali, Landsberg et al. 2012). Sections were imaged using a Tecnai F30 operating at 300kV, equipped with a LC1100 lens-coupled CCD camera (Direct Electron). A 3D dual-axis tomogram was recorded using SerialEM and reconstructed using the IMOD software package (Kremer, Mastronarde et al. 1996).

3.2.1.2 Vitrified mitochondria:

To evaluate the performance of the RAZA filter on vitrified membranes and membrane proteins, a dataset of vitrified mitochondria isolated from *P. anserina* (Davies, Strauss et al. 2011) and imaged by cryo-ET was used. Young *P. anserina* mitochondria were isolated from liquid cultures inoculated with 6 day old mycelium and purified. Sections were imaged using an FEI Polara electron microscope equipped with a field emission gun operated at 300 kV. Images were recorded with a 2k × 2k CCD camera fitted with an 863 GIF Tridiem energy filter (Gatan, Pleasanton, CA, USA) at 10–15 µm underfocus at a specimen temperature of 82 K. More details on culturing, purification and Cryo-ET are described elsewhere by Davies, et al (Davies, Strauss et al. 2011).

3.2.1.3 Tomogram of a 2D crystal of the V_oV_1 rotary ATPase:

A 3D tomographic reconstruction of a 2D crystal of the V_oV_1 rotary ATPase isolated from *Thermus thermophilus* membranes as described previously (Gerle, Tani et al. 2006), was kindly provided by Dr. Christoph Gerle (Picobiology Institute, Department of Life Science, Graduate School of Life Science, University of Hyogo, Kamigori, Japan) and Christopher P. Arthur (FEI Company, Portland, Oregon, USA).

2D crystals were grown from mixed micelles of detergent stabilised V_oV_1 complexes and synthetic lipids (POPC) as described previously (Tani, Arthur et al. 2013). 2D arrays were negatively stained with 2% uranyl acetate on continuous carbon grids with gold fiducial markers (15 nm) applied to the back of the carbon film. A dual-axis tilt series of 101 tilted images (1° increment over a range of ± 55°) was collected at 40,000 × magnification (corresponding to a pixel size of 6.1 Å) using a Tecnai F30 microscope (FEI, Eindhoven, Netherlands) equipped with a post-column energy filter (GIF) and a 2K × 2K CCD camera (Gatan, Pleasanton, CA). 3D reconstruction was performed following the weighted back projection method in the IMOD tomography package (Kremer, Mastronarde et al. 1996, Mastronarde 2005).

3.2.1.4 Cryo tomogram of *B. bacteriovorus* HD100 cells:

Cryo-tomograms of *B. bacteriovorus* HD100 cells kindly provided by Dr Yi-Wei, Jenson lab, California Institute of Technology, US. *B. bacteriovorus* HD100 cells were grown as described

previously (Lambert and Sockett 2005) on *E. coli* S17-1 prey cells in Ca-HEPES buffer at 29°C until most of the prey cells were cleared from the culture. 10 nm colloidal gold (Sigma-Aldrich, St. Louis, MO) pre-treated with bovine serum albumin was added to the cells to serve as fiducial markers during tomogram reconstruction. 3 µl of the resulting sample was pipetted onto a freshly glow-discharged Quantifoil copper R2/2 200 EM grid (Quantifoil Micro Tools GmbH, Jena, Germany) and plunge-frozen in a liquid ethane propane mixture (with 63:37 proportion) using an FEI Vitrobot mark-III (FEI Company, Hillsboro, OR). The frozen grid was then imaged in an FEI Titan Krios 300 keV field emission transmission electron microscope (FEI Company, Hillsboro, OR) equipped with a Gatan energy filter (Gatan, Pleasanton, CA) and a Gatan K2 Summit direct detector (Gatan, Pleasanton, CA) at the Howard Hughes Medical Institute Janelia Research Campus. Energy-filtered tilt series of images of the cell were collected automatically from −65° to +65° at 1° intervals using the UCSF Tomography data collection software (Zheng, Keszthelyi et al. 2007) with total dosage of 100 $e^-/Å^2$, a defocus of -8 µm and a pixel size of 4.2 Å. The images were aligned and subsequently reconstructed into a tomogram by the weighted back-projection method using the IMOD software package (Kremer, Mastronarde et al. 1996).

3.2.2 The RAZA algorithm

The RAZA algorithm is a two-step filter that first applies a Gaussian filter to denoise the image, followed by Laplacian filtration to detect edges. A threshold is then applied to the whole 3D edge-detected volume based on arbitrarily assigned Z-crossings values (ranges from negative to positive or vice-versa). The end result of the filter is a binary volume highlighting all the edges detected throughout the whole tomogram. As electron tomograms are inherently noisy, the prior application of Gaussian filtering reduces the sensitivity of the Laplacian filter to noise. Applying the Laplacian filter directly to a noise corrupted input tomogram can yield many small spurious edges that detract from the larger more meaningful edges with higher SNR. To overcome this, the RAZA filter has two parameters that can be tuned: sigma (σ, the width of the Gaussian which controls the degree of smoothing) and the Z-crossings value (a threshold that defines which edges are detected). Typically the user will first broadly examine different parameter combinations, before fine-tuning to determine optimal parameter settings for the task at hand. Thus, it provides flexibility to users and the ability to test multiple combinations of parameters. The detailed process flow of the RAZA algorithm is summarized in Figure 3.1. This shows that in step 1, RAZA applies a Laplacian of Gaussian (LoG) filter to suppress noise, thereby effectively smoothing the image to yield a LoG-filtered volume for subsequent edge detection. In step 2 the Z-crossings algorithm is applied to the LoG-filtered image to identify the second derivative, which is *the rate of change of the gradient of intensity* (i.e. Z=0 or Z>0 or Z<0), as an edge. Note that sigma represents the radius of the Gaussian

kernel, which optimizes the Z-crossings value (i.e. negative and positive Z crossing values) to define edges. Step two then yields the binary volume.

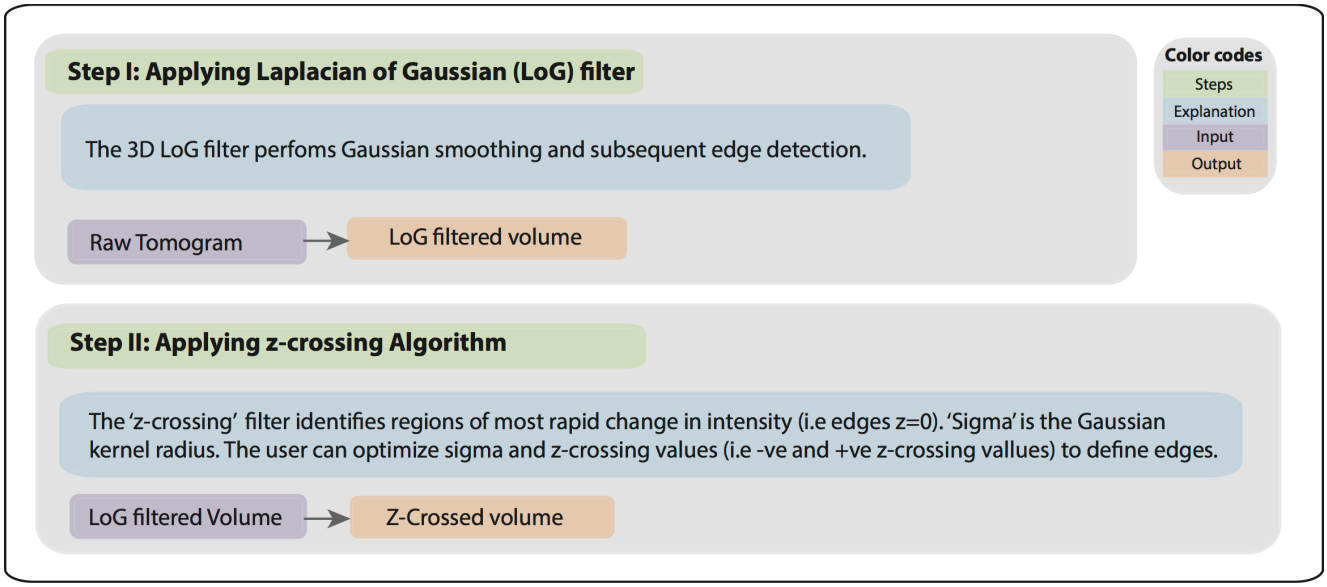


Figure 3.1 Flow diagram representation of various steps involved in the RAZA.

3.2.2.1 3D Laplacian of Gaussian (LoG) method

The first step of the RAZA algorithm requires the generation of a Laplacian of Gaussian (LoG) volume. The Gaussian function (Rafael C. Gonzalez 2002) smooths the focal voxel based on the sigma value which defines the radius of its 3D kernel. In the second step, the Laplacian filter (Rafael C. Gonzalez 2002), then highlights regions with significant changes in voxel intensity.
In three dimensions, the Gaussian function of a volume with continuous voxels (x, y, z) is defined in Equation 3.1:

$$G(x,y,z) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2+z^2)}{2\sigma^2}} \quad (3.1)$$

where x, y and z are coordinates of the focal voxel and σ determines the radius of the Gaussian (G) kernel. Similarly, the Laplacian (L) function derives the second derivative of a 3D object, which is defined in Equation 3.2:

$$L(x,y,z) = \frac{\partial^2 I(x,y,z)}{\partial x^2} + \frac{\partial^2 I(x,y,z)}{\partial y^2} + \frac{\partial^2 I(x,y,z)}{\partial z^2} \quad (3.2)$$

where, $L(x,y,z)$ refers to the Laplacian function or second derivative of the input volume. $L(x,y,z)$ is defined as the sum of the second derivative of the test volume in terms of (x, y, z) respectively.

Finally by applying the Laplacian, $LoG(x, y, z)$ onto a Gaussian-filtered volume (see Equation 3.3);

$$LoG(x, y, z) = C\left(\frac{x^2}{\sigma_x^4} - \frac{1}{\sigma_x^2} + \frac{y^2}{\sigma_y^4} - \frac{1}{\sigma_y^2} + \frac{z^2}{\sigma_z^4} - \frac{1}{\sigma_z^2}\right) e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} - \frac{z^2}{2\sigma_z^2}} \quad (3.3)$$

where C is an arbitrary, real constant (set to 1 in the implemented code). It is of note that the use of small values of sigma results in a reduction in the Gaussian kernel radius. This is best suited to preserving fine details but does not give the same degree of smoothing as a larger sigma value, which can assist with the detection of larger objects. The $LoG(x, y, z)$ operation yields calculated voxel intensities, which range from negative to positive or vice-versa.

3.2.2.2 Arbitrary Z-crossing and traditional zero-crossings concept in 3D

Mathematically, an edge is defined as a region of increased discontinuity. A zero crossing occurs at a point where the sign of the function changes from positive to negative or negative to positive and defines a region of the most rapid change in discontinuity, which can be used to define an edge or a boundary in an automated manner. In the zero-crossings scenario where the output $LoG(x, y, z)$ volume contains an intensity value equal to '0', it marks this as a point on an edge.

To determine whether these potential edge points belong to a contour or not, each voxel is scanned to define whether:

(1) $LoG(x, y, z)$ response 'Z' is zero, positive or negative in the areas of constant intensity of an edge, darker side (one side of an edge) and lighter side (other side of an edge) respectively.

(2) If its neighbouring voxels in 3D (i.e. 26-way connectivity of voxels) have a Z value of opposite sign.

If the above conditions are fulfilled the voxel will be considered a true edge voxel. This process will always generate a closed contour. Theoretically however, the contour defining an edge does not have to be traced along the position of the greatest rate of change (i.e. Z = 0). The selection of a non-zero value, either positive or negative, allows the user to fine tune the position of the traced edge as part of this semi-automated process of tomogram segmentation. Furthermore while the use of z=0 detects all edges in the tomogram, the selection of a non-zero arbitrary Z value can assist with the detection of specific objects.

The selection of a non-zero value, either positive or negative, allows the user to fine tune the position of the traced edge as part of this semi-automated process of tomogram segmentation.

Furthermore while the use of z=0 detects all edges in the tomogram, the selection of a non-zero arbitrary Z value can assist with the detection of specific objects.

During an arbitrary Z-crossing run, the program checks whether the focal voxel intensities are greater or less than an arbitrary number 'Z'. That is, if the Z value is negative then a contour is seeded by a LoG voxel (26-way connectivity) less than or equal to Z.

3.2.2.3 Implementation

RAZA was developed as an integrated plug-in for IMOD (Kremer, Mastronarde et al. 1996) and uses its image processing libraries (Kremer, Mastronarde et al. 1996). RAZA was encoded in C and the source code has been compiled and tested on Mac OSX operating systems (Snow Leopard and Lion). An automatic IMOD script packaged with the software provides the user with an alternative to complex command line operation. RAZA thus does not require that the user has a detailed knowledge of image processing algorithms.

3.3 Results

The RAZA software was designed to provide structural biologists with a rapid, accurate, robust and automated approach for high throughput detection of edges that define organelles, macromolecular assemblies and membrane proteins in cellular tomograms. As an example of organelle segmentation, Figure 3.2A shows a 3D view of an extracted sub-tomogram of a high-pressure freeze-substituted, plastic-embedded mouse islet cell. The outer mitochondrial membranes (red arrow) and inner cristae (blue arrows) are clearly visible, as are other structural features surrounding the organelle. Figure 3.2B shows a representation of the tomogram following Gaussian filtering and subsequent calculation of the first derivative *(rate of change of intensity)*. The *X* and *Y*-axes of Figure 3.2B represent the X-Y tomogram axes while the *Z*-axis represents the range of first derivative values colour-coded from blue (0) to red (2000). The highest 1st derivative values (i.e. the peaks uppermost in Figure 3.2B) represent voxels exhibiting the highest rate of intensity change; i.e. voxels centered around an edge. As 1st derivate filters do not discriminate between signal and noise they cannot accurately distinguish between *true* and *false* edges. In contrast Figure 3.2C represents the Laplacian of Gaussian (LoG i.e. second derivative of Gaussian) of panel Figure 3.2A. As in Figures 2A and 2B, the *X* and *Y*-axes of Figure 3.2C represent those of the tomogram. Here however, the Z-axis indicates the 2nd derivative values (*the rate of change of the gradient of intensity)*. Closer analysis shows that the Z-crossings values can range from positive (red) to negative (blue). The zero crossings point is indicated in yellow. In this dataset positive and negative Z values correspond to the dark and light (grey) background in Figure 3.2A respectively. The real space tomogram (Figure 3.2D) has been rendered by tracing the zero crossings value calculated in

Figure 3.2C, to show edges representing the greatest rate of change in intensity. The automated application of the Z crossings program to Figure 3.2A yielded a segmented tomogram, which included all contours detected including those of the mitochondria and internal cristae. The unwanted contours were manually deleted as shown in Figure 3.2C. While deletion of unwanted contours is necessarily manual, the amount of time required to do this represents a marked improvement upon the full manual segmentation of the tomogram.

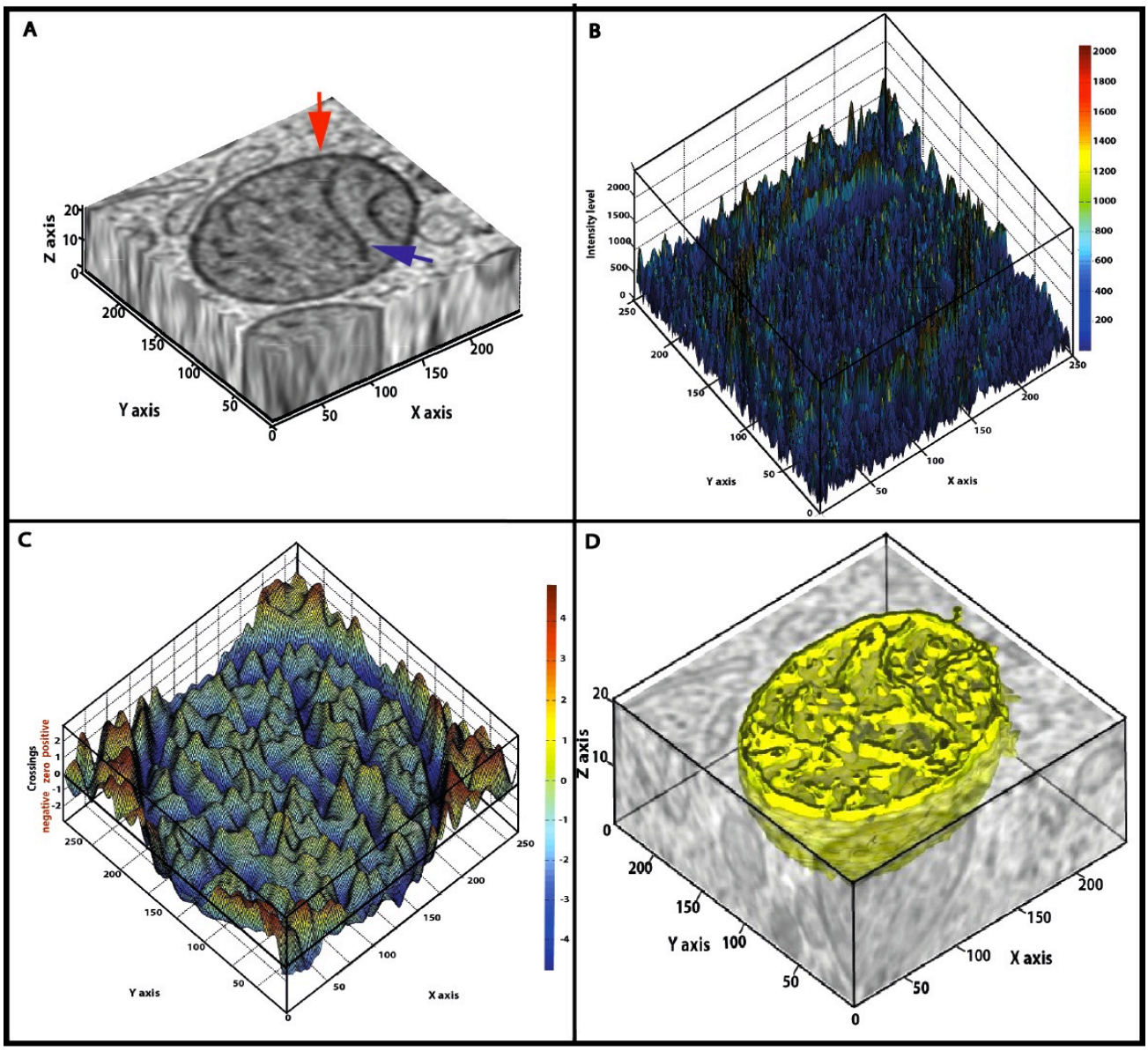


Figure 3.2 The RAZA filter concept.

A) The raw subvolume containing a mitochondrion of an insulin-secreting pancreatic beta cell. B) The 1st order derivative of the Gaussian filtered tomogram. The *X* and *Y*-axes are the *X*-*Y* tomogram axes of panel A. The *Z*-axis represents the range of 1st order derivatives of the greyscale intensity values of Figure 3.2A, which are colour-coded from, blue (0) to red (2000) and peak around the edges. C) The 2nd order derivative of the Gaussian filtered tomogram. The *X* and *Y*-axes are the *X*-*Y* tomogram axes. The *Z*-axis represents the 2nd order derivative values colour coded from positive (red) to negative (blue). Yellow colouring indicates zero crossings. D) The rendered tomogram (Z-crossing value = 0 and σ = 0.49). Z-crossing values were traced for the full tomogram and unwanted contours were manually deleted, leaving the rendered mitochondrion.

3.3.1 Segmentation of a whole cell tomogram

Next, RAZA was tested using a more complex tomographic cellular subvolume acquired for a 300-400 nm high-pressure frozen, freeze-substituted and plastic-embedded mouse islet pancreatic cell section. This volume contained a range of subcellular compartments differing in size, shape, orientation and grey scale values. The automated detection of such a wide range of objects remains challenging, severely limiting the ability to segment and annotate tomographic data in an automated manner. Figure 3.3B shows the Gaussian filtered output of Figure 3.3A, which upon closer analysis is seen to be slightly smoothened. Figure 3.3C shows application of the Laplacian of Gaussian (LoG) followed by zero-crossing to tomogram showed in Figure 3.3A. This image contains the full range of Z-crossings values from the positive to negative range similar to that shown in the colour rendition in Figure 3.2C. Figure 3.3D is an ensemble of edges detected in the slice from Figure 3.3A by applying different Z-crossing thresholds. The yellow contours were obtained at a Z-crossings setting of 1170 using σ = 0.58, the purple contours at z = -1 and σ = 0.49, and the red contours with settings of z = -2405 and σ=0.49. σ can be increased in size to minimize noise, or reduced to preserve fine details and these values were tuned to select the objects of interest. Figure 3.3E shows these same contours (purple, yellow and red) overlaid upon the grey scale image in Figure 3.3A. The yellow contours can be seen to define the small white subcellular structures (e.g. Golgi ribbon- Figure 3.3I), while the purple and red contours delineated the light and dark grey objects in the image (e.g. mitochondria and mature (insulin) granules respectively). Figure 3.3F shows the merged 3D contours for all three sets of objects. (G, H and J) shows 3D surface view of all the white, gray and dark gray objects respectively.

For the purposes of annotation, analysis and selective representation it is important for biologists to have tools to extract specific objects. Figures 3.3G – 3.3I show that through the selection of specific sigma values and Z-crossings thresholds, the white, grey and dark grey objects can be selectively segmented and extracted. These results therefore demonstrate the capability of RAZA to conduct accurate, semi-automated segmentation of a wide range of subcellular volumes in resin-embedded samples, irrespective of their size, shape, volume, orientation and voxel intensity.

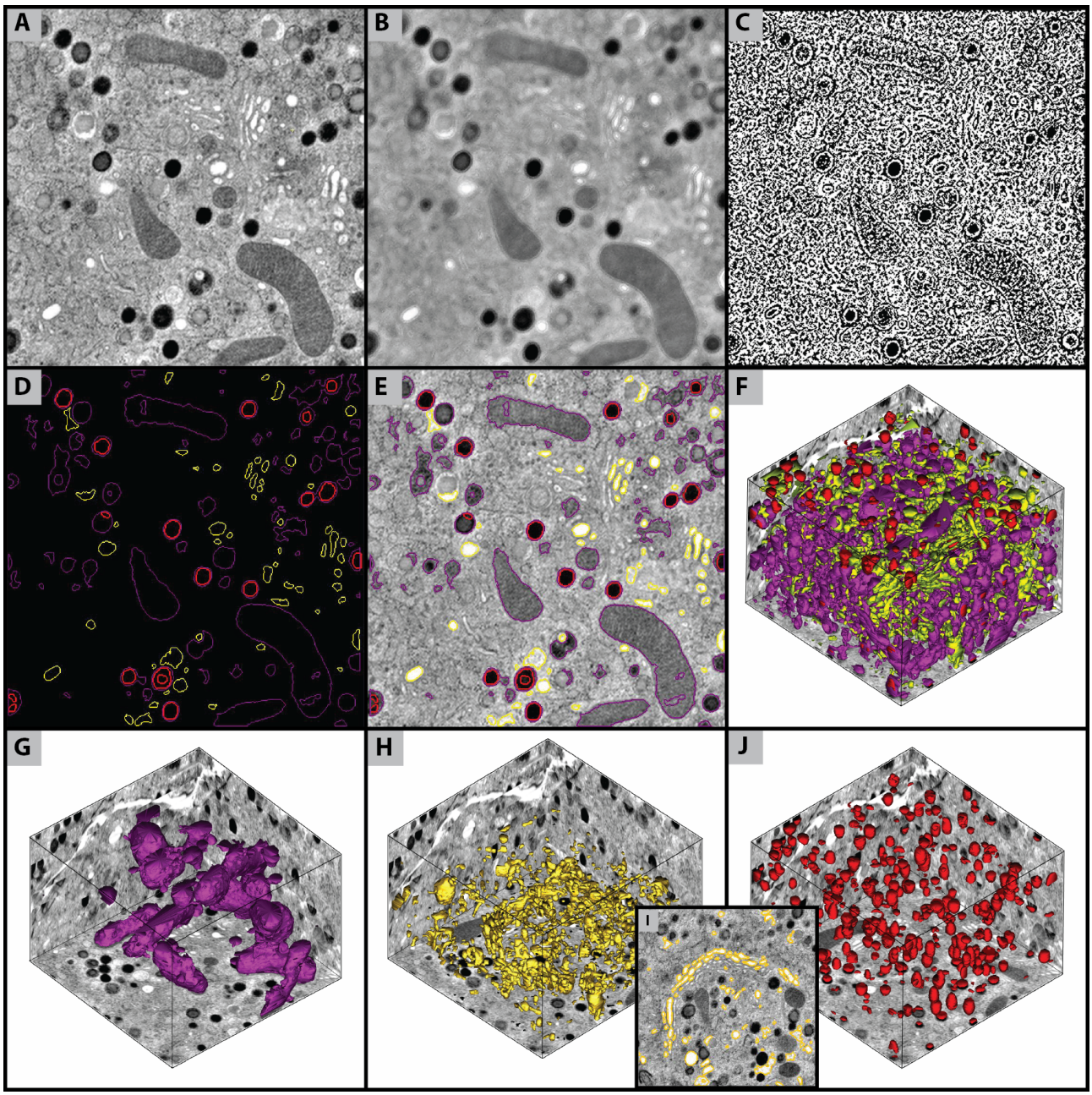


Figure 3.3 Application of the RAZA filter to a whole cell tomogram.

(A) Tomographic volumes that encompassed key compartments involved in insulin production and release by beta cells. (B) Gaussian filtered tomogram shown in A. (C) Application of Laplacian of Gaussian (LoG) followed by zero crossing to tomogram shown in A. (D) The contours detected by RAZA filter: White objects (mainly Golgi regions - a complete and more clear Golgi region with contours detected by RAZA is shown in I) contoured in yellow (Z crossing value = 1170 and σ = 0.58), light gray objects (mainly mitochondria) contoured in purple (Z crossing value = -1 and σ = 0.49), dark gray objects and dark circular objects (mature (insulin) granules) contoured in red (Z crossing value = -2405 and σ= 0.49). (E) All the contours overlayed on the 2D tomographic view shown in A. (F) All the contours represented in the 3D tomographic view. (G, H and J) shows 3D surface view of all the white, gray and dark gray objects respectively.

3.3.2 Application of RAZA to Cryo-tomography.

The segmentation of low contrast cryo-tomograms is substantially more challenging than the segmentation of high dose tomograms recorded on resin-embedded, heavy metal stained specimens. The performance of RAZA applied to cryo-tomographic data is summarised in Figure 3.4 (for comparison, the performance of RAZA on a resin-embedded and negatively stained mitochondrial membrane sub-tomogram is shown in Figure 3.4A-C). Figure 3.4D shows a cryo-tomographic slice of a mitochondrion from *P. anserina*. Manual segmentation yielded Figure 3.4G, where the cristae membranes and several ATP synthase complexes, coloured according to (Davies, Strauss et al. 2011) are clearly resolved. It should be noted that such segmentation requires expert knowledge, and is labour intensive, precluding high throughput analysis. The ability of RAZA to segment membranes and membrane-embedded proteins under low dose cryo-imaging conditions was thus investigated. The tomogram shown in Figure 3.4D was processed using settings of σ=0.49 and z=10. These settings yielded the LoG second derivative image shown in Figure 3.4E from which the segmented 3D volume shown in Figure 3.4F was obtained. These settings were optimized to detect the protrusions seen in the cristae, which correspond to ATPase complexes (see Figure 3.4G). Figures 4H (an enlarged and coloured version of Figure 3.4F) shows a similar distribution of ATPase densities to those segmented manually in Figure 3.4G. The smoothing of the ATPase complexes in Figure 3.4H is due to the fact that here, the settings have been optimised for membrane detection rather than the detection of individual membrane proteins.

The above results show the capability of RAZA to conduct semi-automated segmentation of both membranes and large extracellular domains of membrane-embedded proteins. Although RAZA did not resolve protein contours within the transmembrane region of the ATPase complexes in these images, the ability to localise membrane protein complexes of this size (via identification of their cytoplasmic domain) enables extraction and subsequent subvolume averaging using single particle tomography approaches (Castano-Diez, Kudryashev et al. 2012).

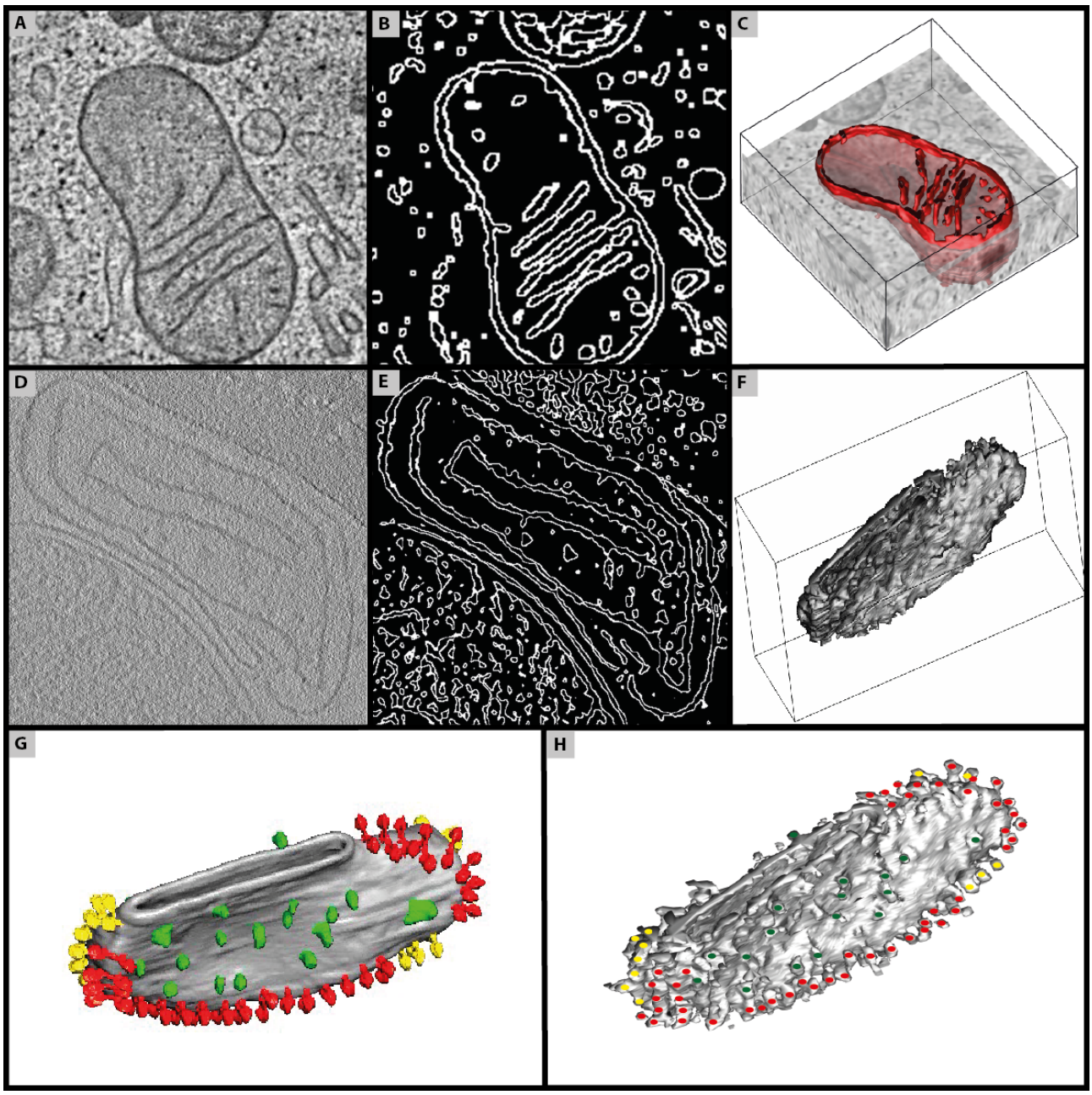


Figure 3.4 Detection of membranes and membrane protein.

(A) Negative stained (fixed) tomographic slice of a mitochondria and its cristae from a pancreatic beta cell. (B) 2nd order derivative of A using RAZA filter settings of z=10 and σ=0.52 showing the detected mitochondrial and cristae boundaries. (C) 3D surface view of segmented mitochondria obtained by RAZA after post-processing to delete unwanted edges. (D) A cryo-tomographic slice containing a membrane (crista vesicle from *P.anserina*). (E) Application of RAZA filter using σ =0.49 and z=10 showing the detected cristae boundaries. (F) 3D surface view of a segmented mitochondrial crista obtained by RAZA. (G) 3D surface of crista membrane. ATPase complexes observed protruding from the membrane are colour-coded as in (Davies, Strauss et al. 2011). (H) An enlarged representation of the 3D surface view shown in F with the detected ATPase complexes colour-coded according to G.

3.3.3 Comparing manual and automated segmentation

Having established that RAZA can segment both resin embedded and cryo-tomograms of biological samples accurately and in a semi-automated manner, studies were conducted to evaluate its performance (Figure. 3.5D-F) in comparison with existing peer-reviewed and published manual

segmentation approaches (Figure 3.5A-C) (Marsh, Mastronarde et al. 2001). Figure 3.5A shows a tomogram of a high pressure frozen, freeze-substituted, resin-embedded pancreatic beta cell, which in Figure 3.5B has been overlaid with a number of manually segmented contours of clathrin-positive endo-lysosomal compartments (red), mitochondria (dark green) and Golgi (pink, red, light green and blue) which differ significantly in size and shape (Marsh, Mastronarde et al. 2001). Figure 3.5C shows the extracted 3D objects defined by these contours. For comparison, Figure 3.5D shows the corresponding contours generated using RAZA settings selected to detect these different objects. These contours are overlaid on the tomogram shown in Figure 5A. Overlaying these contours onto the tomogram (Figure 3.5E) shows that the quality of the output is high and similar to that of manual segmentation (Figure 3.5F). Furthermore a closer comparison of Figure 3.5C and Figure 3.5F shows that the RAZA filter appears to have retained a higher level of structural detail.

Table 3.1 Comparison of surface area and volume.

RAZA and manual segmentation are used to compare surface area and volume. Percentage difference is also indicated.

Object Name	Surface Area (SA) µm^2			Volume (V) µm^3		
	RAZA	Manual	Dev (%)	RAZA	Manual	Dev (%)
Mitochondria	146624	147857	0.84	376389	365226	2.97
Golgi	967716	983916	1.67	2296214	2328761	1.42
Endolysosomes	216835	150638	30.53	403028	267398	33.65

To statistically compare the accuracy of the RAZA filter with manual segmentation, we quantified the surface areas and volumes obtained using both techniques. Specifically we calculated the surface area and volume of mitochondria, Golgi and endolysosomal elements as representatives of differently sized objects generated manually and using RAZA (Table 3.1). The surface area of Golgi elements and mitochondria closely matched using both segmentation methods (percentage difference less than 2%). We also note that the volumes of mitochondria (dark green) and Golgi elements (pink, red, light green and blue**)** obtained using RAZA and manual segmentation differ by a maximum of 2.97%. When we compared the volume and surface area of endolysosomal compartments (red), which are smaller and have a lower SNR than the Golgi and Mitochondria using both methods, we found that they differ by up to 33.65%. This difference is likely due to the fact that RAZA delineated the contours in a more detailed way and also included a few additional

(true positives using manual inspection) objects thereby increasing the total surface area and volume detected. While this may be indicative of a limitation of the RAZA filter, automated segmentation and subsequent editing of segmented objects is nonetheless a marked improvement on manual segmentation.

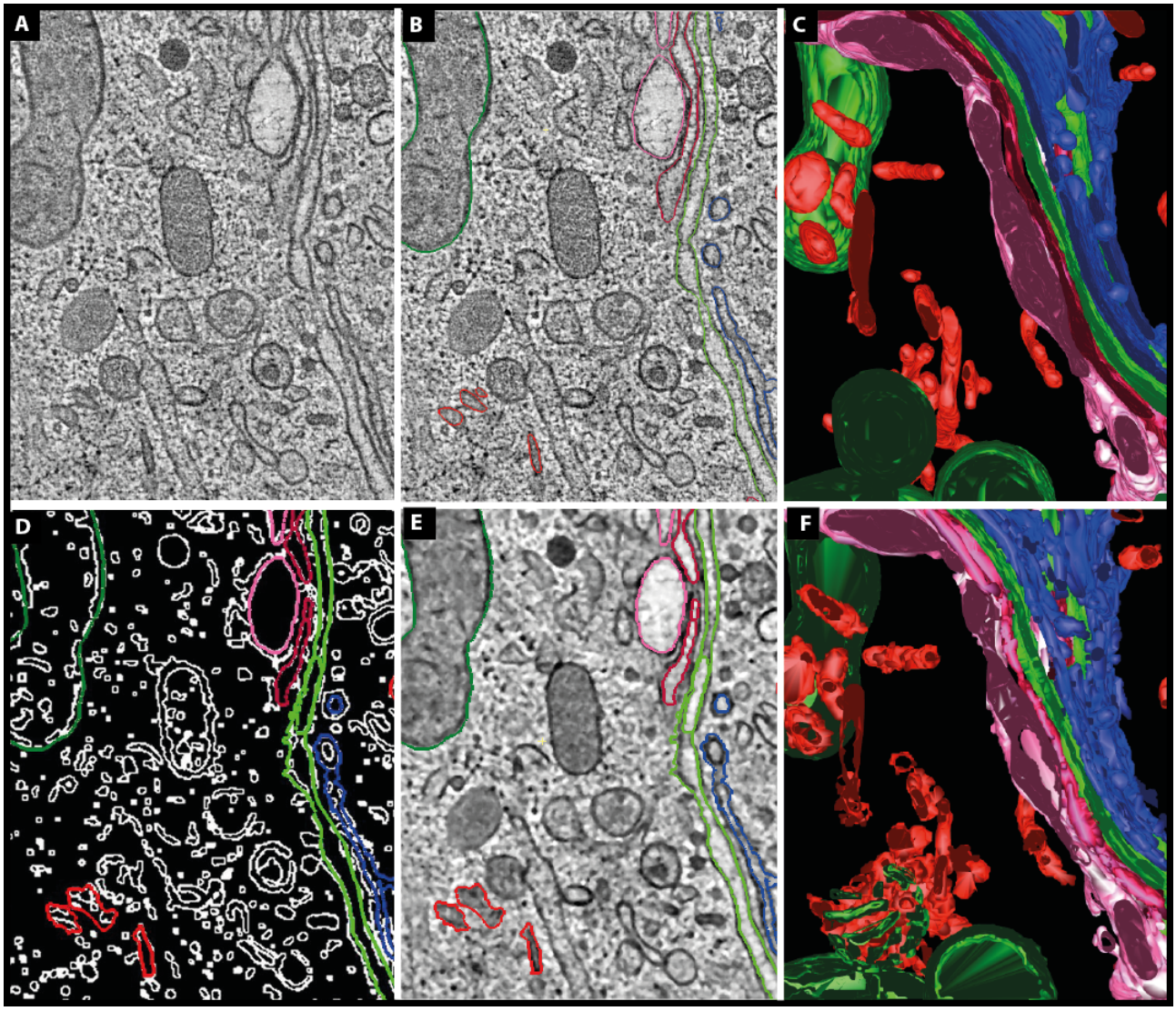


Figure 3.5-Segmentation of the Golgi region of an insulin-secreting pancreatic beta cell.

(A) A slice of the raw tomographic volume. (B) Manually contoured clathrin-positive (red) endolysosomal compartments, mitochondrial (dark green) and Golgi (pink, red, light green and blue) which differ significantly in size and shape (Marsh, Mastronarde et al. 2001). (C) 3D representation of the manually contoured objects. (D) RAZA traces contouring clathrin-positive endolysosomal compartments (red, z=-10 and σ= 0.49), mitochondrial (dark green, , z=2 and σ=0.49) and Golgi (pink, z=-10 and σ=0.49; red, z=-10 and σ=0.49; light green, z=-10 and σ=0.49 and blue, z=-10 and σ=0.49) which are overlaid upon the binary output (z=-2 and σ=0.49) and on actual tomogram in D and E respectively. (F) 3D representation of the RAZA-contoured objects.

3.3.4 Detecting macromolecular assemblies in electron tomograms.

Having established that RAZA compares in quality to best practice manual segmentation at the level of subcellular organelles, its ability to contour individual macromolecular assemblies was

tested next. Figure 3.6A shows a tomogram from a cytoplasmic region of a high-pressure frozen, freeze-substituted, resin-embedded *C. reinhardtii* cell containing a large number of darkly stained macromolecular assemblies of a comparable size (and though to be) ribosomes. Using RAZA settings of z=-1 and σ=1, contours defining the outline of these macromolecular assemblies could be rapidly generated (Figure 3.6B). Figure 3.6C shows a 3D surface view of these contoured macromolecular assemblies with Figure 3.6D showing a higher magnification view of the region demarcated by a red box in 6C. This clearly demonstrates the ability to obtain 3D information defining the spatial arrangement of macromolecular assemblies in cells. Furthermore the ability to localise and rapidly delineate a large number of individual macromolecular assemblies paves the way for the extraction of large 3D datasets for subsequent alignment, classification and molecular subvolume averaging.

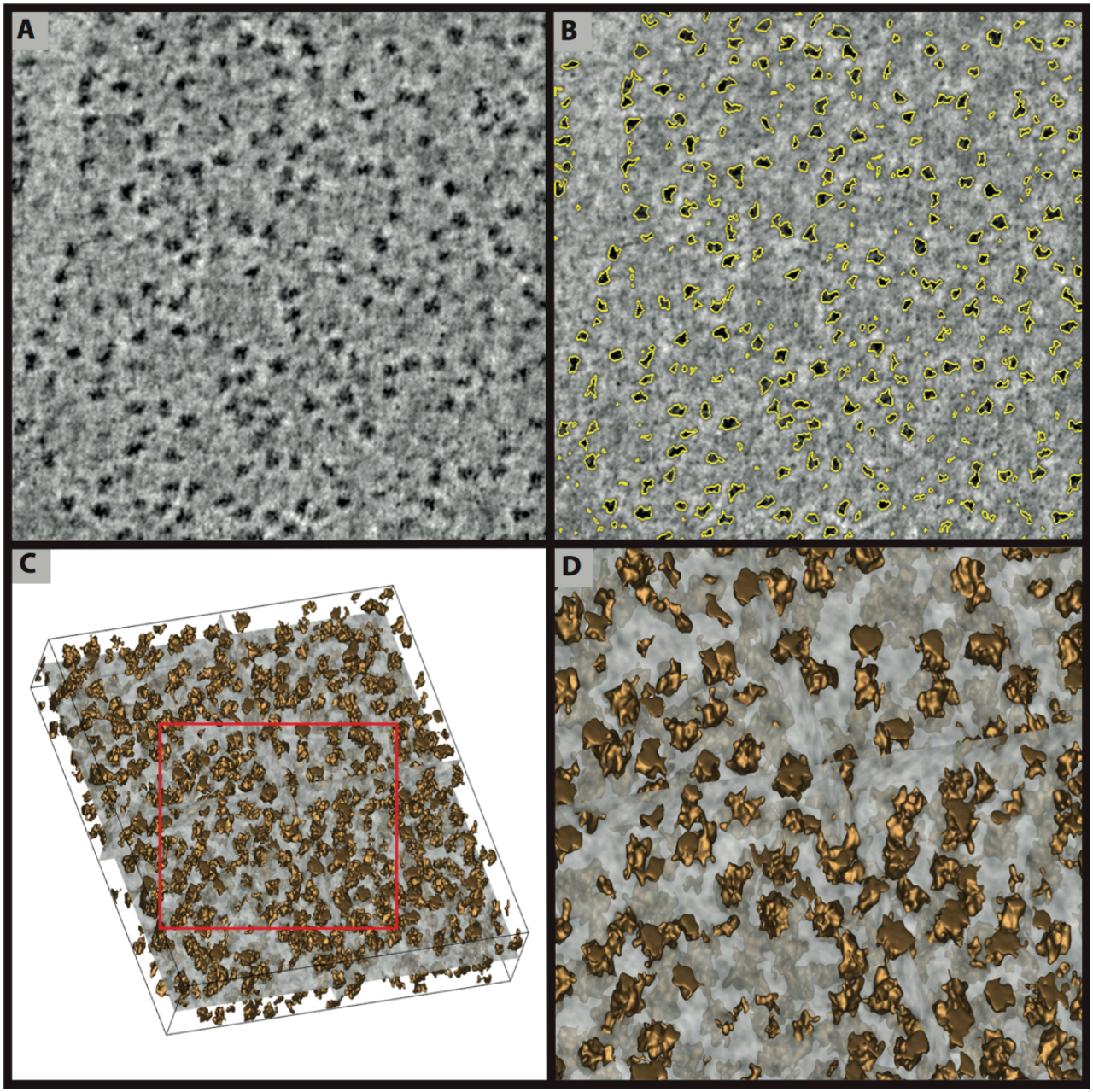


Figure 3.6 Segmenting macromolecular assemblies from an electron tomogram.

(A) 2D cross-section of a sub-volume of a resin-embedded C. reinhardtii cell highlighting the region enriched in macromolecular assemblies thought to be ribosomes. (B) Contoured macromolecular assemblies using RAZA settings of z= -1 and σ=1. (C) 3D surface rendered view of B. The region delineated by a red box is enlarged in (D) to show the quality of the RAZA segmentation and the crowded nature of macromolecular organisation.

3.3.5 Segmentation of V_0V_1 rotary ATPases through a 2D crystal.

The capability of RAZA to visualize extra membranous protrusions of transmembrane proteins within mitochondrial membranes (Figure 3.4 H) using selected settings was tested. The final test that was applied was to establish whether RAZA was capable of being tuned to contour transmembrane regions using a 3D tomographic volume obtained from a negatively stained *Thermus thermophilus* V-ATPase 2D crystal. A small region (Figure 3.7A) showing an area of high coherence was extracted from a larger 2D crystalline array, and the RAZA algorithm was applied with parametric settings of Z = 50 and σ = 0.49 (Figure 3.7B). RAZA accurately detected both the extracellular and transmembrane regions of the V-ATPase (Figure 3.7F) in the reconstituted 2D crystal. The accuracy of detection is evident from comparison of Fourier transforms of the raw tomogram (Figure 3.7D) and the segmented volume (Figure 3.7E), where diffraction spots are seen beyond first order reflections; if RAZA were segmenting noise, the reflections in the Figure 3.7E would be lost.

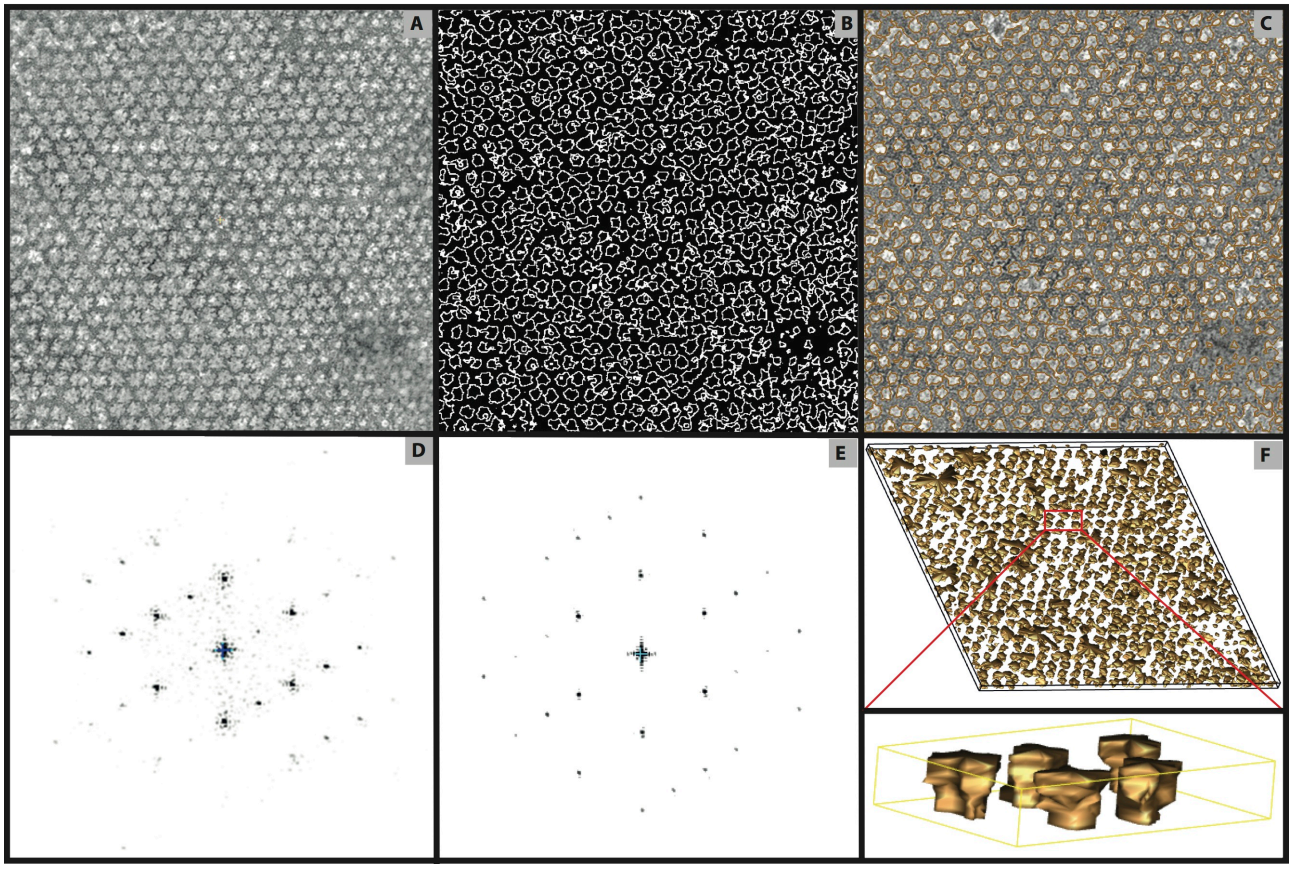


Figure 3.7 Segmentation of V_0V_1 rotary ATPases through a 2D crystal.

(A) 2D cross-section of a sub-volume of a 2D crystalline array of T. thermophilus VoV1 rotary ATPase complexes (Gerle et al., 2006; Tani et al., 2013). (B) Application of RAZA outlining the edges of ATPase complexes shown in (A). (C) Contours (shown in (B) overlaid on the 2D slice shown in (A). These contours/edges were detected using RAZA settings of z = 50 and σ=0.49. (D) FFT of input slice shown in (A). (E) FFT of rendered tomogram shown in (B). (F) 3D surface rendered view of B showing the crowded organisation of V-ATPase complexes within the ordered arrays and the ability of RAZA to segment the transmembrane domains in these reconstituted 2D crystal. Tomograms analysed were kindly provided by Dr. Christoph Gerle (Picobiology Institute, Department of Life Science, Graduate School of Life Science, University of Hyogo, Kamigori, Japan) and Christopher P. Arthur (FEI Company, OR).

3.3.6 Segmentation of membranes in cryo tomogram.

In recent years, there has been a significant increase in the importance of using cryo-ET datasets and especially for the extraction of membranes within a cellular context. To evaluate RAZA using real cryo-tomograms, first the edge detection of the outer membrane obtained from *B. bacteriovorus* cells (Figure 3.8A) was attempted. The σ parameter determines and adjusts the width of the object of interest while the Z value improves the connectivity of true edge pixels by observing the intensity values obtained based on the second order derivative. In this cryo-tomogram, the optimal σ value was found to be 0.7 for both membranes and membrane protrusions, as they both have similar widths. They did however have different optimal Z values of 378 and 360.

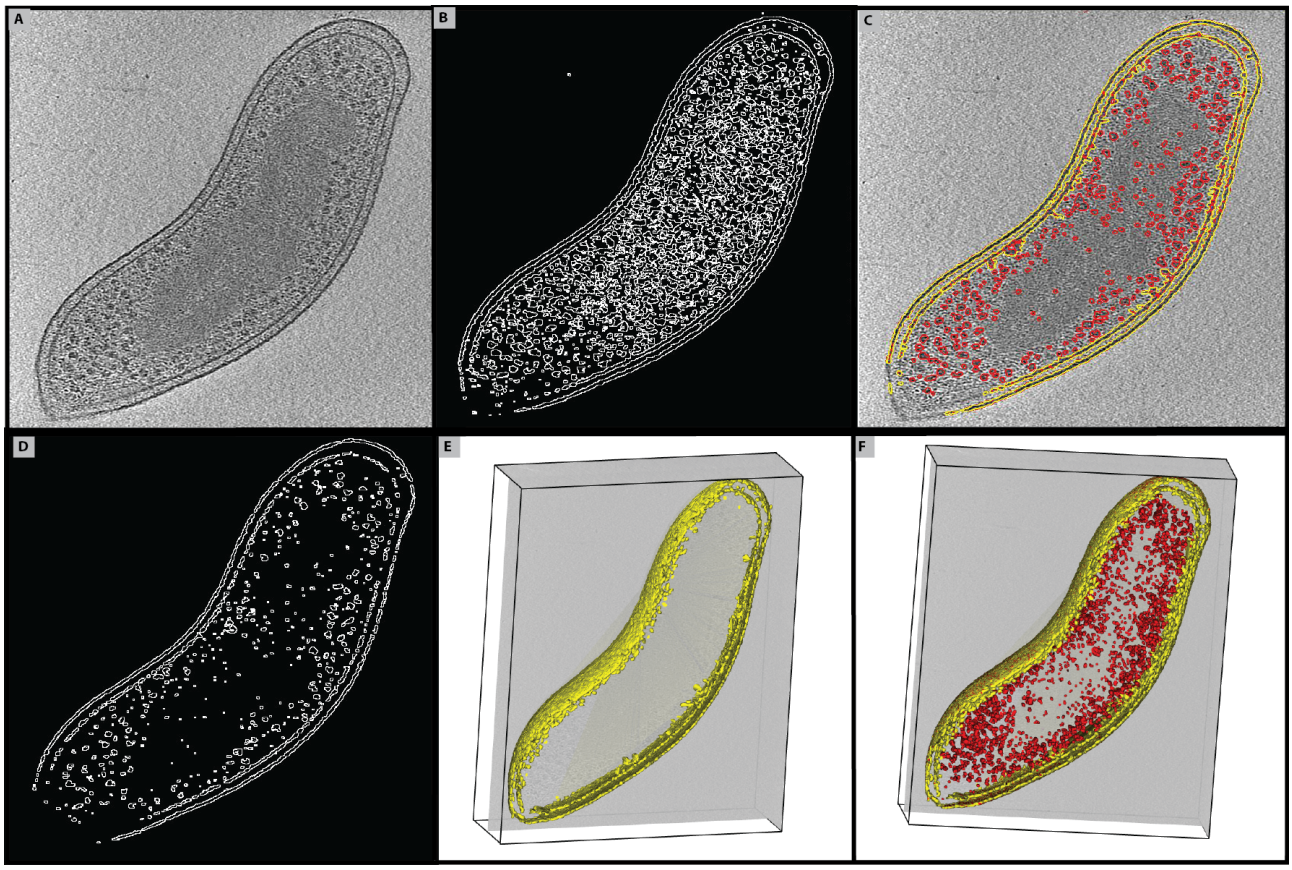


Figure 3.8 Segmentation of membranes of *Bdellovibrio bacteriovorus* imaged under cryo conditions.

(A) A 2D slice of a cryo-tomogram containing a membrane of B bacteriovorus HD100. (B) RAZA segmentation output using σ = 0.7 and Z = 378 optimised for membrane detection (C) Contours detected by RAZA with settings mentioned in B (in yellow) and D (in red) overlaid on the 2D tomographic view shown in A. (D) Application of RAZA using σ = 0.7 and Z = 360 optimised for macromolecular structures. (E) 3D surface view of Membranes only. (F) Combined 3D surface view of membranes and macromolecules detected by RAZA using both parametric settings. Tomograms analysed were kindly provided by Dr Yi-Wei, Jenson lab, California Institute of Technology, US.

To select protrusions within the membranes more accurately a Z value of 360 was used to eliminate false edges (Figure 3.8D). Despite the low signal to noise ratio and other artefacts of cryo-tomography (e.g. missing cone), RAZA recovered contours (Figure 3.8 E and F) close to those expected for the membrane (Figure 3.8A).

3.4 Discussion and Conclusion

The results presented here show that RAZA is able to conduct rapid, selective and completely automated segmentation of resin-embedded, negatively stained cellular tomograms, as well as cryo tomograms. RAZA can also be applied to a broad range of other 3D imaging technologies (e.g. MRI, CAT) as the process is based on the generic principles of Gaussian noise suppression and edge detection. Specifically RAZA defines *the rate of change of the gradient of intensity* (i.e. Z=0 or Z>0 or Z<0) thereby offering the flexibility to adjust the arbitrary Z-crossings value right across the positive to negative spectrum of the image, enabling the selective segmentation of objects (e.g. differing in greyscale values). Adjusting σ allows the user to find an appropriate balance between noise suppression and the retention of high-resolution structural detail. Despite its simplicity, RAZA has proven effective at segmenting a broad range of objects under a wide range of conditions. Another important aspect of this generic approach is the fact that detectable edges span a broad, multi-scale range.

In this case the range covers volumes as large as the outer contour of the cell down to the level of edges that define individual membrane proteins complexes in membranes (Figure 3.4 and Figure 3.7) and macromolecular assemblies such as ribosomes (Figure 3.6). Despite its simplicity, RAZA has therefore proven effective at segmenting a broad range of objects including organelles, macromolecular assemblies (e.g. ribosomes) and membrane proteins in membranes (e.g. ATPases) under a wide range of conditions demonstrating its effective operation across a broad, multi-scale range.

RAZA generates continuous 3D contours each of which effectively defines a discrete geometric volume with quantifiable properties (e.g. length, width, height, 3D volume and surface area) which can theoretically be used both for object detection (e.g. the selection of macromolecular assemblies in a particular size range) and potentially, automated cellular annotation and molecular docking approaches. For example, an object with one very long axis and two short axes might represent a

tubule, while macromolecular assemblies will have a surface area in a range significantly smaller than most organelles. Furthermore the automated detection of large numbers of macromolecular assemblies within a specific size range offers a significant advantage for high throughput extraction, annotation and downstream single particle tomography for example via Dynamo (Castano-Diez, Kudryashev et al. 2012). In Chapter 4, I aim to present a novel particle-picking tool RAZA^{PS} that uses the continuous edges of RAZA to extract the geometrical properties of segmented objects such as length of major, middle, minor axes, surface area, volume and center of segmented object to further facilitate the particle selection.

Collectively, the ability to automate the detection of edges across multiple scales and to produce contours which are mathematically defined opens up significant potential not only to examine the spatial relationship of membrane proteins, macromolecular assemblies and organelles within the cellular context, but to provide a structural framework for the docking of high resolution protein structures obtained using crystallography, NMR and advanced single particle analysis. In so doing, RAZA enables multi-scale structural analysis and has the potential to provide important capability on the path to developing pseudo-atomic 3D reconstructions of e.g. subcellular and organellar structures through the integration of tomographic, SPA and atomic resolution structural data. Consequently RAZA provides an important new capability to bridge the micron-nanometer resolution range as part of an overall strategy to develop pseudo-atomic resolution atlases of cells.

CHAPTER 4

RAZA^{PS}: Rapid, Automated Z-Crossing Algorithm for 3D particle selection.

As emphasized in preceding chapters that revealing the 3D molecular architecture of cells including the *in situ* organization of organelles, macromolecular assemblies and membrane proteins is at the forefront of structural biology. In Chapter 2 and 3, we attempted to address the fundamental research question to develop an accurate 3D edge detector and benchmark it against available methods. We benchmarked 3D BLE against 3D Canny and Recursive filters in Chapter 2 whereas Chapter 3 shows that RAZA is better suited for detecting fully connected macromolecules as compared to 3D BLE. Specifically, in Chapter 3, we showed that the edge detection process facilitated by RAZA helps in segmenting objects in both resin-embedded and cryo-electron tomograms. Accurate segmentation of macromolecules in electron tomogram can help develop structural fingerprints by defining each object mathematically to further facilitate object quantification and selection. In this chapter, I address one of the major research goals of this PhD, "to use edge detection methods to develop 3D particle-picking tools for the quantification of macromolecular assemblies and subcellular compartments by defining *structural fingerprints* using geometric properties of a segmented object." I also aim to use these structural fingerprints to facilitate detection of ribosomes in *C. reinhardtii*. Furthermore, I aim to use the tools developed in this thesis to understand the changes in the distribution, size and number of mitochondria.

Chapter 4 therefore presents an advanced *Rapid, Automated Z-crossing algorithm* for *Particle Selection* (RAZA^{PS}) to automate tomogram segmentation and particle selection for further sub-tomogram analysis. It is a high-throughput Laplacian-of-Gaussian edge-detector followed by arbitrary Z-crossing, which outputs mathematically defined object contours. Each object has a longest axis (*major axis*), a second longest axis orthogonal to this (*middle axis*) and the third, mutually orthogonal axis to these (*minor axis*) as well as a calculated surface area and volume. These 5 parameters are object identifiers referred as *structural fingerprints*. Calculating these at the

object level overcomes the problem of the 'random' orientation of subcellular structures. RAZA^{PS} can segment all detected objects, or a selected population of objects (e.g. mitochondria or ribosomes). The specificity of object selection is fine-tuned by the Gaussian radius (δ), the z-crossing value and the *structural fingerprint* parameters. Biological diversity is accounted for by tuning individual threshold settings for each of the 5 structural fingerprint parameters (e.g. 70-130% of the major, middle and minor axes, surface area and/or volume). The number of particles selected is most sensitive to the threshold setting of single parameters (e.g. length). We show that by using all 5 *structural fingerprint* parameters, we achieve maximum automated selection of related objects. RAZA^{PS} calculates object centers for high-throughput sub-volume extraction and averaging. It also enables the construction of 'atomic resolution' 3D models of cells as the molecular contours and their object identifiers provide the structural framework for the semi-automated docking of atomic protein structures into cellular tomograms.

4.1 Introduction

Electron microscopy is currently undergoing a 'resolution revolution' (Kuhlbrandt 2014) due to a number of fundamentally important advances at the level of imaging, data capture and image processing. Collectively these are resolving the ultrastructure of cells and their constituent protein complexes ever more quickly and to higher resolution. For example, cryo-focused ion beam (Cryo-FIB) milling is now making it possible to resolve the ultrastructure of whole cells to ~4nm resolution in vitreous ice (Engel, Schaffer et al. 2015). The integration of phase plates (Danev, Buijsse et al. 2014) into advanced electron microscopes is increasing the contrast of cellular densities recoverable by cryo-transmission electron microscopy into the sub-nanometer range, thereby improving the visualisation of macromolecular assemblies and ordered subcellular structures in a near-native environment (Asano, Fukuda et al. 2015). Their resolution in turn is being enhanced further through the use of sub-volume averaging programs (e.g. Dynamo (Castano-Diez, Kudryashev et al. 2012), PyTom (Hrabe, Chen et al. 2012). Furthermore state-of-the-art direct electron detectors greatly reduce the impacts of sample charging and drift on the captured image, increasingly making atomic resolution structure determination by single particle analysis (Allegretti, Mills et al. 2014) a reality not only for viruses and macromolecular assemblies but also for membrane proteins smaller than 500kDa (Li, Mooney et al. 2013, Liao, Cao et al. 2013, Lyumkis, Julien et al. 2013, Bartesaghi, Matthies et al. 2014, Lu, Bai et al. 2014, Campbell, Veesler et al. 2015).

Based on these advances, structural biology is now entering the exciting era of resolving atomic resolution 3D atlases of cells based on the nested docking of multiscale datasets (from whole cell tomograms to atomic resolution protein structures). Tomograms, upon segmentation, provide

structural constraints to dock atomic resolution structures of large macromolecular assemblies and protein complexes into cellular tomograms (Davies, Strauss et al. 2011). Accurate, high-throughput cell segmentation capability is therefore a critical link between advanced electron tomography and both high-throughput sub volume averaging and precise molecular docking.

While manual tomographic segmentation processes are well established (Marsh, Mastronarde et al. 2001) and automated routines are being developed and implemented (mostly operating as 2D algorithms), automated segmentation processes capable of organelle and macromolecular segmentation in 3D are less advanced. Ideally such automated processes must be: *(1) Robust, (2) Accurate, (3) Simple to use, (4) Fast, (5) Generally applicable*: i.e. capable of operating in a comprehensive mode to select all sub-volumes including organelles and macromolecular assemblies, and *(6) Selective*: i.e. capable of operating in a selective mode to isolate defined target structures (e.g. only ribosomes).

Improvements in resolution assist in this process. This however is counter-balanced by the fact that as spatial resolution increases, there is a cubic increase in the volume of data captured increases as a function of x^3, making the automation of image denoising and segmentation ever more important for the recovery of high resolution detail at a meaningful rate.

Edge detection and segmentation algorithms operating either at the 2D or 3D level have been designed around a number of principles including Laplacian of Gaussian (Woolford, Hankamer et al. 2007), fast marching methods in which a contour is initialised from a pre-chosen seed point and allowed to grow until a certain boundary condition is reached (Bajaj, Yu et al. 2003, Baker, Yu et al. 2006), bilateral edge detection (Ali, Landsberg et al. 2012), Eigen-vector analysis (Frangakis and Hegerl 2002), watershed transforms (Volkmann 2002), optimal thresholding (Cyrklaff, Risco et al. 2005), orientation fields based segmentation of membranes (Sandberg and Brega 2007), and fuzzy set theory principles (Garduno, Wong-Barnum et al. 2008).

Fortunately, atomic models are already available for many proteins (PDB). These atomic coordinates can be used to generate electron density maps (Forster, Han et al. 2010). Jochen Bohm and his colleagues were among the first to demonstrate detection of macromolecules (thermosomes, 20S proteasome and *E. coli* GroEL) within real tomographic data using cross correlation-based template matching (Bohm, Frangakis et al. 2000). Two years later, Achilleas Frangakis published a fully automated template-matching algorithm, which was able to detect and identify crowded 20S proteasomes and thermosomes, in a tomogram of phantom cells (Frangakis, Bohm et al. 2002). For more objective annotation and contouring of microtubules and ER-Golgi membrane-like cellular structures in electron tomogram, template matching (Lebbink, Geerts et al. 2007, Lebbink, van Donselaar et al. 2009) has proven to be very useful. For example, Weber and colleagues (Weber,

Greenan et al. 2012) proposed an automated segmentation of electron tomograms based on a 3D template matching algorithm that enables automated tracing of microtubules and provides a quantitative description of actin filament .

Template matching has been demonstrated to be effective for tomogram segmentation of specific and uniform target objects such as ribosomes. However as its name suggests, template matching requires a specific template to detect specific target objects. This approach is not ideal for high throughput and automated segmentation of the large diversity of organelles, macromolecules and membrane proteins in the cell, as many templates would be required even to approach full segmentation. Nor is the method suited for the identification of many types of organelles, as these tend to have diverse structures. Additionally template-matching algorithm usually yield a considerable number of false positives for noisy datasets. For example the so called 'Einstein from noise' effect is amongst the drawbacks of template matching methods; that is the cross-correlation of noise can result in a false positive giving the illusion of correctly contoured object. Shatsky et al., (Shatsky, Hall et al. 2009) and Henderson R (Henderson 2013) discussed the false positives produced by model-based template matching. Shatsky et al. (Shatsky, Hall et al. 2009) performed the alignment and averaging of 1000 images with pure white noise and reconstructed *Einstein* image (that was provided as a template) from white noise images. Furthermore van Heel M (van Heel 2013) demonstrated the ability to generate a false trimeric HIV-1 envelope glycoproteins structure from random noise to highlight the pitfall of template matching methods.

In contrast particle picking based on edge detection algorithms offer the advantage of being able to segment objects (i.e. defined by an edge) in an automated manner. So for robust high throughput analyses RAZA^{PS} is a preferred approach for particle selection and tomogram as it detects edges regardless of the objects size, shape or orientation.

However, most of the above mentioned algorithms have only been implemented in 2D or pseudo 3D and thus have limited utility for analyzing the data contained within 3D image volumes in a truly 3D context. A true 3D filter, capable of using data from adjacent slices, offers the advantage that additional information from either side of the 'focal' slice can be considered, thereby enabling enhanced noise suppression along with the detection of contiguous and legitimate structural details throughout the 3D image stack. Also, these edge detectors often yield non-continuous edges, which may not be perfectly suited to subsequent analyses.

One of these segmentation algorithms is the 'Rapid, Automated 3D Z-Crossings Algorithm' RAZA (Chapter 3). It is based upon the use of an adjustable *Gaussian filter* to de-noise the image, and *Laplacian of Gaussian* analysis to enable the calculation of the second order derivatives which

measures the *rate of change of the gradient pixel intensity*, for edge detection. The ability to measure the *rate of change of the gradient of pixel intensity* (i.e., Laplacian) increases the ability to distinguish between real and spurious edges and so improves detection results (See Chapter 3).

We have shown in Chapter 3, the RAZA filter has the advantage that it is *robust* (applicable to any 3D image volume), *accurate* to the level of detecting macromolecular assemblies in cells and extrinsic domains of membrane proteins in membranes. It is *simple to use* (2 parameters i.e., sigma and Z-crossings value), *fast* (estimated to be 4000 times faster than manual segmentation) and can be operated either in a *general mode* to segment all detected objects or in a *selective* manner to identify specific objects within in a defined target range. Moreover, RAZA provides continuous edges to facilitate accurate *structural fingerprints*. Because of all these advantages over other algorithms, it was selected to provide a strong base for further particle picking analysis.

An integral part of RAZA's functionality is that each of the contoured objects that it generates can be defined as a single mathematical entity (not just a set of contouring pixels), having a measurable *length*, *width*, *height*, *surface area* and *volume*. These five parameters essentially provide a '*structural fingerprint*' for each segmented object, which can be used to selectively identify discrete classes of subcellular objects. Furthermore these parameters are independently calculated for each object, thereby overcoming the analytical problem of the random orientation of organelles and protein complexes in cells. This is illustrated in Figure 4.1, which shows a cell volume (pink) containing tubules (yellow, long and thin), organelles (blue, rounded) and macromolecules (orange). In this example, the small macromolecules and can be readily distinguished from the larger organelles and tubules as they have significantly smaller volumes, surface areas and dimensions defined by the major (longest), middle (second longest axis orthogonal to the major axis) and minor axis (orthogonal to the major and middle axis). The rounded organelles (blue) and tubules (yellow) which may or may not have similar volumes and surface areas can be readily distinguished as the latter have one very long axis (major) and two much smaller axes (middle and minor). In contrast the three axes of the organelles (blue) are much more similar in length. Natural variability of organelles can also be addressed through the use of threshold setting for each of these five parameters (e.g. +/-30% of a given reference value). Furthermore by reducing the search range threshold (e.g. +/-5% of a given reference value) and/or increasing the number of the five variables used for the search (e.g. all versus volume only) the stringency with which particles are identified can be significantly increased.

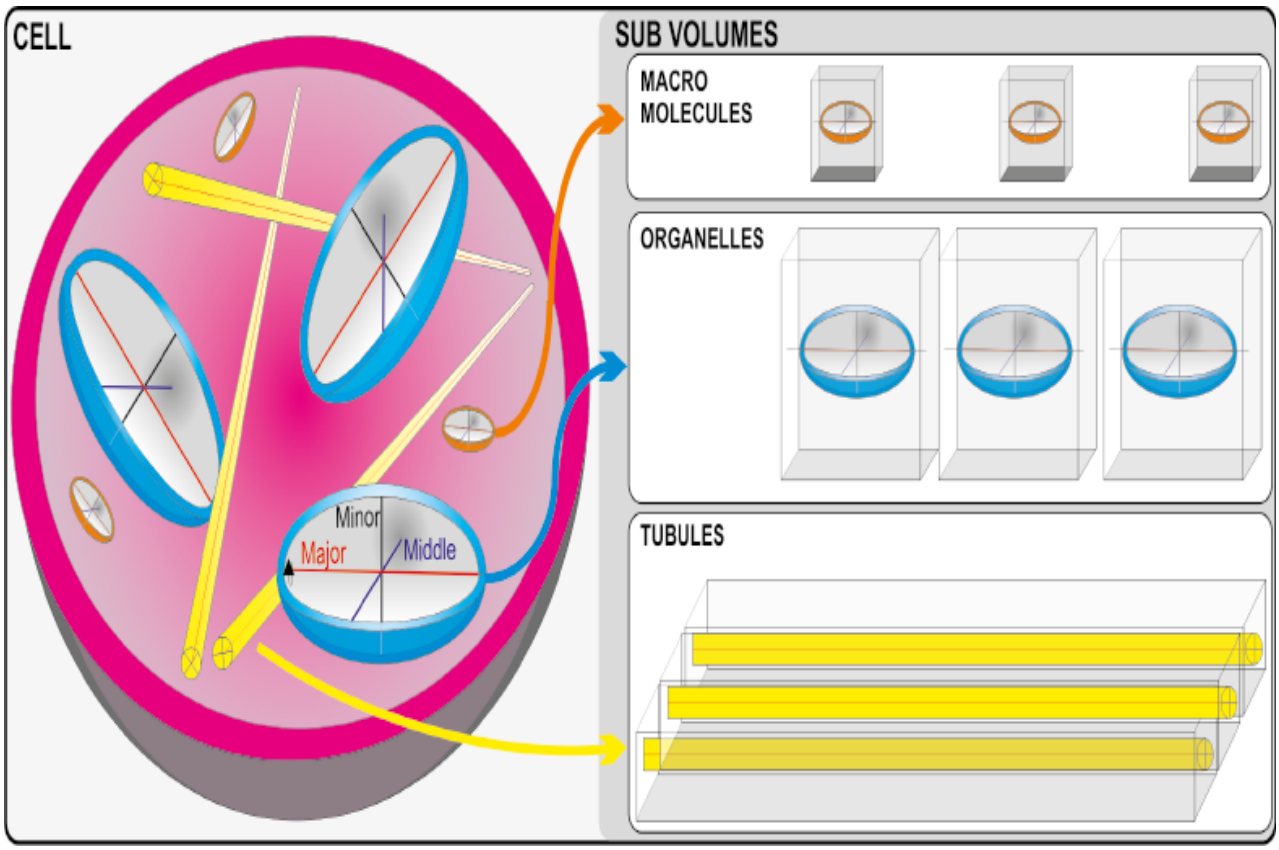


Figure 4.1 Depiction of RAZA^{PS} method.

RAZA^{PS} based structural fingerprinting for selective high throughput segmentation, object identification and sub-volume extraction. Each object is a mathematically defined unit having a discrete volume, surface area, major (longest axis), middle (second longest axis orthogonal to the major axis) and minor axis (third axis orthogonal to the major and middle axis). These parameters act as *structural fingerprints* enabling selective particle detection and are independent of the random orientations of organelles and protein complexes in cells. Object centers are also calculated to enable sub-volume extraction.

This approach opens up the possibility of high-throughput '*general*' and '*selective'* cellular segmentation, the extraction of large numbers of tomogram sub-volumes of a defined target object (e.g. ribosomes) and potentially, coarse alignment based on these parameter sets. Furthermore, this approach provides a bridge between the ever increasing amount and quality of electron tomography data and the development of high throughput processes for sub-volume averaging, structural population analysis (e.g. classification of organelle populations based on size and shape) and the development of atomic resolution 3D atlases of cells by docking atomic models of proteins into the contours of macromolecular assemblies in tomograms.

4.2 Methods

4.2.1 Electron Tomographic Datasets

4.2.1.1 GroEL truth set:

The performance of RAZA^{PS} was initially tested using a 'truth set' volume containing 100 copies of the GroEL chaperonin complex (EMDB accession code 1081) having different orientations (Ludtke, Chen et al. 2004, Ali, Landsberg et al. 2012). This was designed to define the statistical variation within the detected surface areas, 3D volumes, and lengths of the major, middle and minor axes.

4.2.1.2 Truth set containing mixed objects:

The accuracy and integrity of the RAZA^{PS} was evaluated using a truth set comprising conceptual reference volumes representing different geometries (20 copies of GroEL chaperonin complex (EMDB accession code 1081) having different orientations, 64 small sized spheres, 28 large sized spheres). The volumes were designed to thoroughly test the discriminative power of RAZA^{PS} to accurately determine the true positive rates, false positive rates, true negative rates, false negative rates, 'receiver operator characteristic' curves (ROC) and 'area under' ROC (AUROC). The test volume contained randomly oriented GroEL, as well as the small and large spheres and was contaminated with 60 % Gaussian noise (Figure 4.8).

4.2.1.3 High-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell tomogram:

To evaluate the performance of RAZA^{PS} features on tomographic data (Tecnai F30, 4700x magnification), a whole cell tomogram reconstructed from a high-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell isolated from Islets of Langerhans were used (Noske, Costin et al. 2008). This dataset provides a rigorous test as it contains subcellular structures defined by a broad range of grayscale values ranging from close to white (e.g. Golgi), light gray (e.g. mitochondria) to dark gray/black organelles (e.g. insulin secretory granules) as well as object dimensions, surface areas and volumes.

4.2.1.4 Chloroplast tomogram:

A chloroplast region of a high pressure frozen, freeze substituted plastic-embedded *C. reinhardtii* cell (Ali, Landsberg et al. 2012) was imaged using a Tecnai F30 electron microscope operating at 300kV (magnification = 23000x), using a lens-coupled CCD camera (Direct Electron LC1100). A 3D dual-axis tomogram was recorded using SerialEM and reconstructed using the IMOD software

package (Kremer, Mastronarde et al. 1996). This dataset contains a multitude of ribosome-like macromolecules.

4.2.2 RAZA^{PS} overview - from tomograms to sub-tomograms

RAZA^{PS} has been developed as an integrated plug-in for IMOD (Kremer, Mastronarde et al. 1996) and so is freely available for academic distribution. The IMOD environment was chosen as it is widely used for the 3D reconstruction and image analysis of cellular tomograms and provides a powerful framework for the development of new interfaces for different 3D algorithms. RAZA^{PS} has been encoded in the C programming language. We successfully compiled the source code for RAZA^{PS} and tested it on Mac OSX operating systems (Snow Leopard and Lion). The user can fine-tune parameter combinations (sigma and Z-crossings value) to identify optimal values for the task at hand. An input file *'UserReferenceModel'* containing the x, y, z coordinates of 'reference objects' is provided to RAZA^{PS}. It then calculates the lengths of the longest axis (*major axis*), then the next longest axis (*middle axis*) orthogonal to the major axis and finally the third longest axis (*minor axis*) mutually orthogonal to both longest and middle axes, as well as the surface area and volume of the reference volumes and stores these in *'ReferenceStructuralDetails'*. Based on these values, a tolerance threshold range can be defined (e.g. -/+ 30% of the 100% reference values). Upon adjusting sigma (Gaussian filter radius) and Z-crossings values (a threshold value ranging from negative to positive or vice-versa that defines which edges are detected), RAZA^{PS} is able to calculate final output sub-tomograms. Two output files are produced: 1) the *'FinalSubvolumeCoordinates'* file which contains the x, y, z coordinates of the centers of the final sub volumes of detected objects; and 2) the *'FinalStructuralDetails'* file which contains all the structural information of detected objects i.e. estimated values of height, width, length, surface area and volume. The program structure provides users with a generally applicable approach and the flexibility to optimize parameters further for a given task, if required. The RAZA^{PS} workflow consists of seven steps and is summarized in Figure 4.2.

4.2.2.1 3D Laplacian of Gaussian (LoG) method

RAZA^{PS} applies a Laplacian of Gaussian (LoG) filter to suppress noise, thereby effectively smoothing the image to yield a LoG filtered volume for subsequent edge detection (Figure 4.2- Step I). This algorithm has two parameters, which affect edge detection: **sigma and the Z-crossings** value. RAZA^{PS} requires the generation of a Laplacian of Gaussian (LoG) volume. First, the Gaussian function (Rafael C. Gonzalez 2002) smooths the focal voxel based on the sigma value which defines the radius of its 3D kernel. In the second step, the Laplacian filter (Rafael C. Gonzalez 2002), then highlights regions of significant change in voxel intensities. In three

dimensions, the Gaussian function of a volume with continuous voxels (x, y, z) is defined in Equation 4.1:

$$G(x, y, z) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2+z^2)}{2\sigma^2}} \quad (4.1)$$

where (x, y, z) are coordinates of the focal voxel and σ determines the radius of the Gaussian (G) kernel. Similarly, the Laplacian (L) function derives the second derivative of a 3D object, which is defined in Equation 4.2:

$$L(x, y, z) = \frac{\partial^2 I(x,y,z)}{\partial x^2} + \frac{\partial^2 I(x,y,z)}{\partial y^2} + \frac{\partial^2 I(x,y,z)}{\partial z^2} \quad (4.2)$$

where, $L(x, y, z)$ refers to the Laplacian function or second derivative of the input volume. $L(x, y, z)$ is defined as the sum of the second derivative of the test volume in terms of *x*, *y* and *z*, respectively. Finally by applying the Laplacian ($LoG(x, y, z)$) onto a Gaussian-filtered volume (see Equation 4.3);

$$LoG(x, y, z) = C\left(\frac{x^2}{\sigma_x^4} - \frac{1}{\sigma_x^2} + \frac{y^2}{\sigma_y^4} - \frac{1}{\sigma_y^2} + \frac{z^2}{\sigma_z^4} - \frac{1}{\sigma_z^2}\right) e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} - \frac{z^2}{2\sigma_z^2}} \quad (4.3)$$

where C is an arbitrary real constant (set to 1 in the implemented code). It is of note that the use of small values of sigma results in a reduction in the Gaussian kernel radius. This is best suited to preserving fine details but does not give the same degree of smoothing as a larger sigma value, which can assist with the detection of larger objects. The $LoG(x, y, z)$operation yields calculated voxel intensities, which range from negative to positive or vice-versa. This step outputs a volume named *'LoG-filtered volume'*

4.2.2.2 Arbitrary Z-crossing vs traditional zero-crossings concept in 3D

Mathematically, an edge is defined as a region of increased discontinuity. A zero-crossing occurs at a point where the sign of the function changes from positive to negative or negative to positive and defines a region of the most rapid change in discontinuity, which can be used to define an edge or a boundary in an automated manner. In the zero-crossings scenario where the output $LoG(x, y, z)$ volume contains an intensity value equal to '0', it marks this as a point on an edge.

To determine whether these potential edge points belong to a contour or not, each voxel is scanned to define whether:

(1) $LoG(x, y, z)$ response 'Z' is zero, positive or negative in the areas of constant intensity of an edge, darker side (one side of an edge) and lighter side (other side of an edge) respectively.
(2) If its neighbouring voxels in 3D (i.e. 26-way connectivity of voxels) have a Z value of opposite sign.

If the above conditions are fulfilled the voxel will be considered a true edge voxel. This process will always generate a closed contour. Theoretically however, the contour defining an edge does not have to be traced along the position of the greatest rate of change (i.e. Z = 0). The selection of a non-zero value, either positive or negative, allows the user to fine tune the position of the traced edge as part of this semi-automated process of tomogram segmentation. Furthermore while the use of z=0 detects all edges in the tomogram, the selection of a non-zero arbitrary Z value can assist with the detection of specific objects.

In implementing an arbitrary Z-crossing approach, RAZA^{PS} checks whether the focal voxel intensities are greater or less than an arbitrary number 'Z'. That is, if the Z value is negative then a contour is seeded by a LoG voxel (26-way connectivity) less than or equal to Z. The 'LoG filtered volume' is then thresholded according to the specified Z-crossings value, to identify regions of most rapid change in intensity (for example Z=0), which correspond to edges (Figure 4.2- Step II). Typically Z-crossings values define where there is a significant change in intensities and based on these values a 'z-crossed volume' is generated.

4.2.2.3 Connected component labelling approach

RAZA^{PS} requires a reference volume, which contains a selected set of objects of interest. To generate a '*UserReferenceModel'* the user can open the tomogram in IMOD and select the object(s) to be used as the reference input model(s) for RAZA^{PS} by using the IMOD 'drawing tool'. The next step drives connected component labelling. After applying the Z-crossings approach (Figure 4.2- Step II) to a raw tomogram, binary volumes containing edges are identified. To distinguish the boundaries of all objects and detect the connected regions of each object in a binary digital volume, a connected component labelling algorithm (Shapiro 2002) was used. Specifically, all connected voxels are labelled and stored as discrete objects, so that identifiable structures are in a sequence that allows the corresponding contour models to be generated as single geometric objects. Each labelled object represents a connected set of contour pixels. The algorithm scans the binary volumes voxel by voxel from top to bottom and left to right (by moving along a x- and y-axes) through multiple passes. In the first pass, it scans through the whole volume, checking the neighbouring voxels, and if the neighbouring voxels are not assigned with a label, it assigns a new unique label, otherwise the focal voxel is assigned the label of the neighbouring voxel. Then, in the second pass it

re-scans all the labels and replaces each temporary label with the unique label of its equivalent class. The pseudo code for connected component labelling is described below

Pseudo Code: Connected component labelling

```
Input: Labeling a particular voxel (x, y, z)
if the voxel (x, y, z) = '0' then
     Do nothing and proceed to next voxel (x+1,y, z)
Else if the voxel (x-1, y, z) has a label then
     Assign the label to the voxel (x, y, z)
Else if neither voxels (x-1, y, z) or (x, y-1,z) is labeled then
     Increment label numbering and assign the latest label to
voxel (x, y, z)
Else if voxels (x-1, y, z) or (x, y-1, z) is labeled then
     Assign the label to the voxel (x, y, z)
Else if both voxels (x-1, y, z) and (x, y-1, z) are labeled then
     Assign the label of voxel (x-1, y, z) to the voxel (x, y, z)
Record the equivalence if labels of voxels (x-1, y, z) and (x, y-
1, z) are not identical.
End if.
```

After this step, each object can then be reduced to a mathematically defined unit having a discrete volume, surface area and major, middle and minor axes each with measurable lengths (Figure 4.2-Step III). The output file of this step is called the *'ComponentLabeledFile'*, which is a 3D volume containing 3D objects labeled as discrete objects.

4.2.2.4 Definition of geometric properties of contoured objects

The structural information of objects in contour spaces used in RAZA^{PS} includes: (1) 3D surface area, (2) 3D volume, (3) height, (4) width, (5) length, and (6) the center of the particle. For **3D surface area,** RAZA^{PS} calculates the length of each contour on a slice (in XY dimension) in a 3D section by adding the number of voxels present within a contour. It then sums the lengths of all the contours in the Z-direction. It also calculates the **3D volume** of the object, based on the total number of voxels within the object.

Next the RAZA^{PS} algorithm calculates the **lengths of the *major*, *middle* and *minor* axes.** It first calculates the length of the longest (*major*) axis of a segmented structure. Thus the *major* axis is defined as the maximum distance between two voxels within an object. The *middle* axis is the

maximum distance between two voxels, with the additional restraint that it be perpendicular to the major axis vector. The *minor* axis is mutually orthogonal to the middle and major axis vectors, bounded by the object contour. These 3D axes provide the height, length and width of the object, independent of object orientation within the tomogram. In this scheme the intersection of the major, middle and minor axes would therefore correspond to the x, y and z coordinates of the object, which is additionally useful for extracting particles for sub volume averaging. Furthermore, the geometric properties of the object can potentially be used to guide initial coarse alignment of extracted particles (e.g. aligning objects along their longest axis). Thus, the parameters extracted by RAZA^{PS} provide a reduced 'structure approximation' for each detected object. For example, the user is able to program that they want to search for and segment objects within defined x, y, z limits e.g. Microtubules have a large x-axis and small y and z axis, whereas for macromolecules, one can select objects less than 50nm in all dimensions.

RAZA^{PS} uses the *'ComponentLabeledFile'* and '*UserReferenceModel'* to completes a 3D volume and 3D surface area calculation (Figure 4.2- Step IV). It also estimates (Figure 4.2- Step V). The length of major axis, middle axis and minor axis of *'reference objects'* are saved in a file *'ReferenceStructuralDetails'*. The same parametric values are calculated from the discrete objects obtained from (Figure 4.2- Step III). An advantage of this approach is that it is independent of the random orientations of objects in cell.

4.2.2.5 Automatic object detection and center determination

RAZA^{PS} conducts semi-automated object detection, calculates their centers and uses the parametric values (width, height, length, surface area and volume) as search parameters to identify objects within the defined structural range. The values of the length of *major*, *middle* and *minor* axes of each object from *'ReferenceStructuralDetails'* are then compared (Figure 4.2- Step VI) with the parametric values obtained from '*ComponentLabeledFile'*. The tomogram is then exhaustively searched to complete automatic particle selection. The details of each identified object is saved in the '*FinalStructuralDetails' file*. Based on this comparison, all the objects within a set tolerance threshold range (e.g. +/- 30%) of the 100% threshold settings of objects selected in the '*UserReferenceModel'* will be detected. After detection of all the objects, RAZA^{PS} provides the capability to box out the detected objects, based on the calculated center of the particles. The segmented tomograms will only show the selected objects.

The identification of the object center enables RAZA^{PS} to conduct automated sub-volume selection and the extraction of centered particles for 3D sub-volume averaging. Once all the necessary

information has been calculated and similar objects have been identified, he next step is to determine $\overrightarrow{\text{MinMax}_x}$, $\overrightarrow{\text{MinMax}_y}$ and $\overrightarrow{\text{MinMax}_z}$ vectors for each object. The voxel where these three vectors intersect determines the center of that object. Mathematically, the vector $\overrightarrow{\text{MinMax}_x}$ passes through MINx (minimum x-coordinate of object on x-axis) and MAXx (maximum x-coordinate of object on x-axis) $\overrightarrow{\text{MinMax}_y}$, passes through MINy (minimum y-coordinate of object on y-axis) and MAXy (maximum y-coordinate of object on y-axis) and $\overrightarrow{\text{MinMax}_z}$ passes through MINz (minimum z-coordinate of object on z-axis), MAXz (maximum z-coordinate of object on z-axis). By using above information the coordinates of the center of a detected object enables rapid sub-volume extraction (Figure 4.2- Step VII). Specifically, the (x, y, z) coordinates of a detected object are used to box out all the detected objects in the form of separate 3D volumes. These files can then be used for alignment and to calculate structures using sub-volume averaging techniques. It is of note that the center of the objects is located at the center of the x, y and z coordinates of the 3D box. In this way step VII achieves the first alignment step for subsequent sub-volume averaging. The major, middle and minor axes provide additional variables for first pass alignment of extracted sub-volumes.

4.2.2.6 Automatic Particle extraction

After detecting all particles of interest within a volume, the last step is to box them out and save them as separate files. To complete this task RAZA^{PS} reads the coordinates of MINx and MAXx, MINy and MAXy and MINz, MAXz from the text file (FinalSubvolumeCoordinates, Figure 4.2; Step VI) and trims the volumes based on user defined volume limits. The extracted sub-volumes are saved as separate 3D volumes. Also note that we do not expect the reference and detected objects to have absolutely same and exact structural fingerprints, therefore RAZA^{PS} gives user option to provide a tolerance threshold for each parameter.

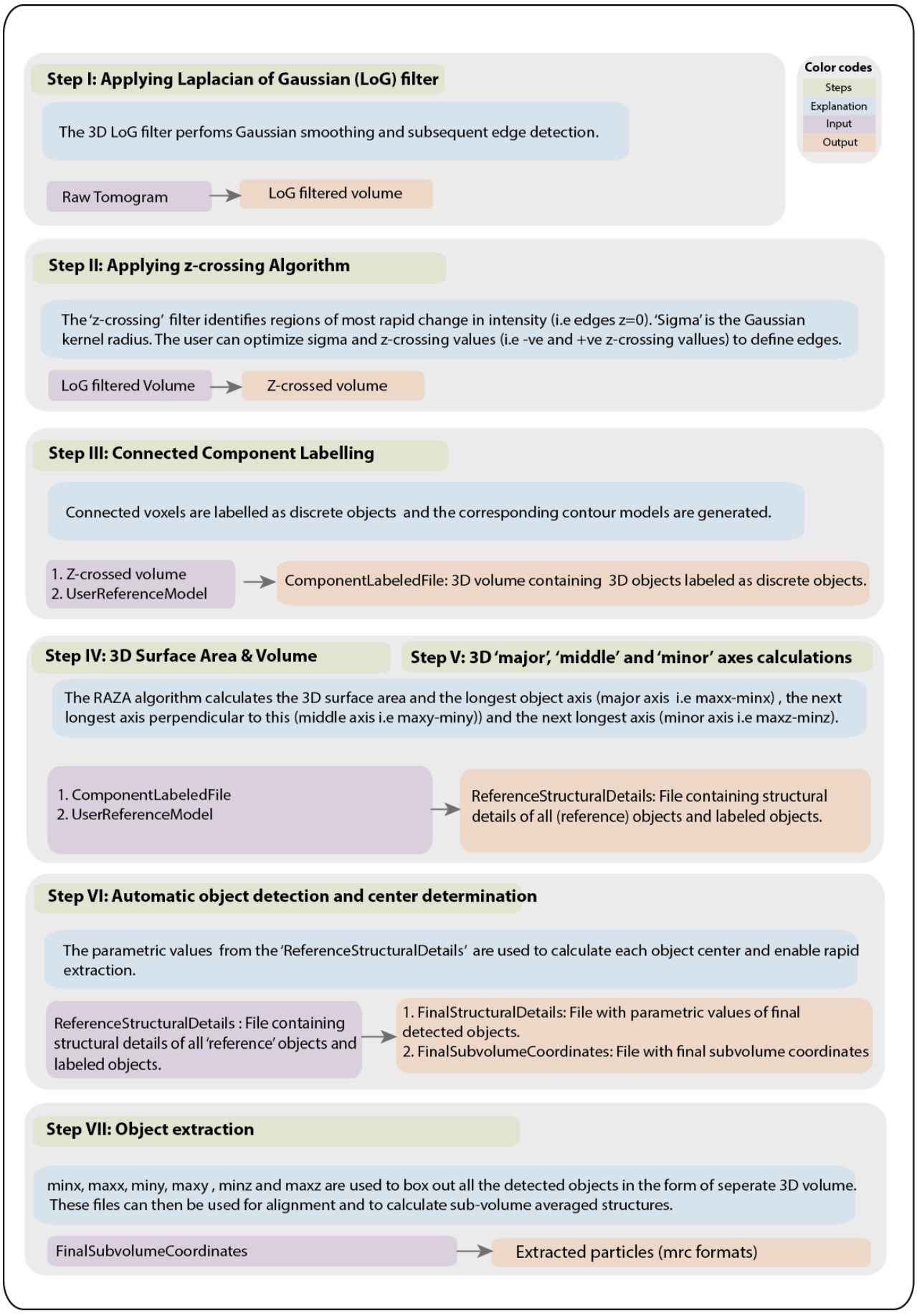


Figure 4.2 Flow diagram representation of various steps involved in RAZA^{PS} algorithm.

4.3 Results and discussion

4.3.1 Accuracy of RAZA^{PS} in a simulated dataset

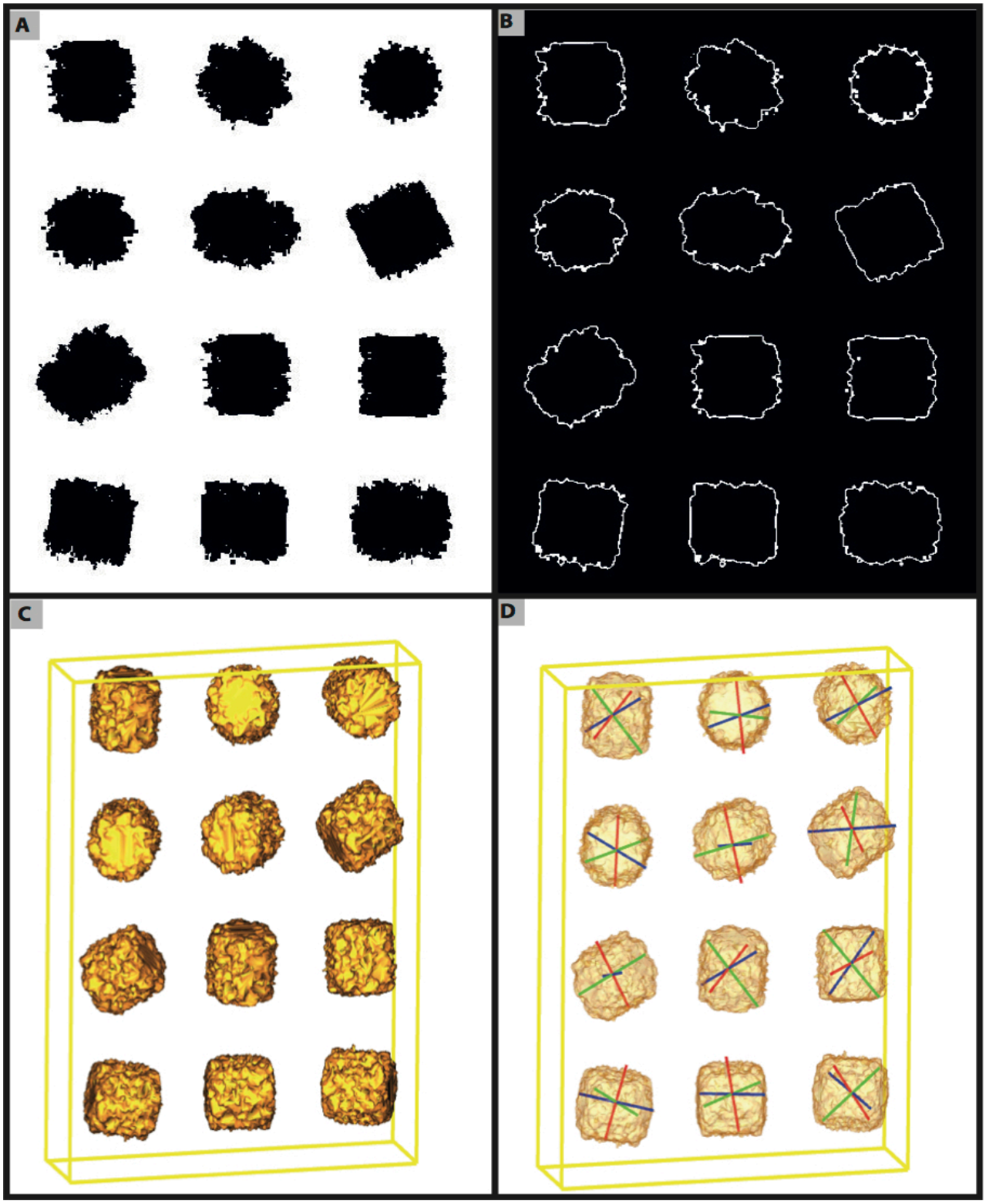


Figure 4.3 Application of RAZA^{PS} to molecular volumes.

The ability of the software to resolve molecular contours was assessed using a test volume populated with 12 GroEL molecules. A) shows the slice taken from 3D volume containing different orientations of synthetic GroEL. B) Application of RAZAPS (Z-crossing value = -1 and σ = 1) which traced outlines around boundaries of the objects. C) Shows the surfaces rendered from detected objects based on all 3 axes (major,

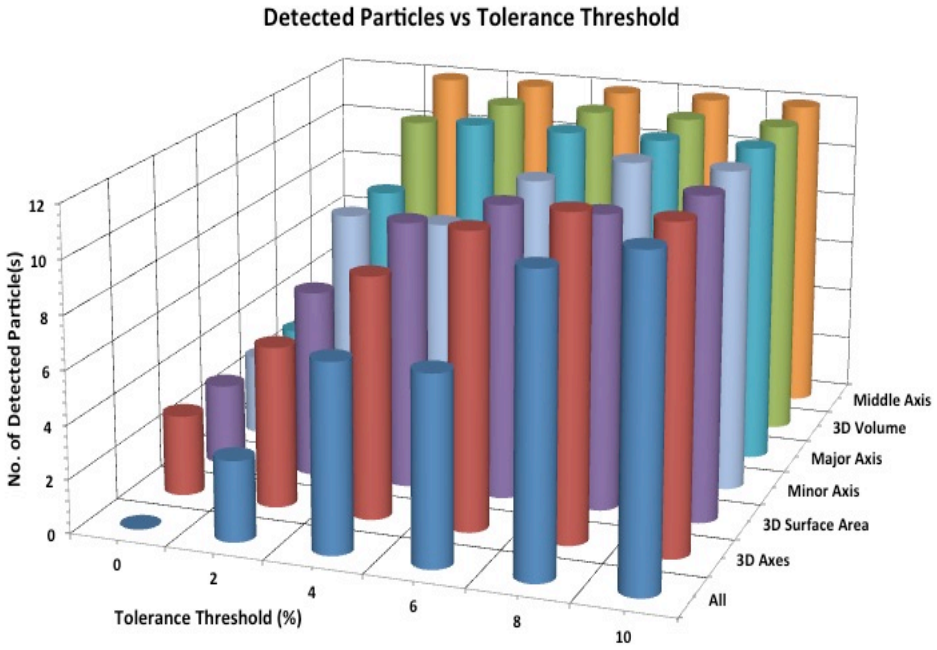
middle and minor), 3D surface area and 3D volume. All these objects were detected at 2% tolerance threshold (by comparing only length of middle axis). D). shows the same detected objects shown in C, but highlighting major (blue), middle (green) and minor (red) all together intersecting on the same point which represent the center of the object.

This chapter aims to provide structural biologists with a rapid, accurate and robust automated approach not only for high throughput detection of edges that define organelles, macromolecular assemblies and membrane proteins in cellular tomograms, but also the detailed structural measurements e.g. depth, length and width, 3D surface area and 3D volume. We initially tested RAZA^{PS} on three different 3D volumes, which represent a broad range of typical microscopy datasets.

We first evaluated the features of RAZA^{PS} on a simulated dataset representing a noise-contaminated volume containing 12 GroEL molecules in different orientations (Figure 4.3). A 2D section taken from this synthetic volume populated with 3D GroEL molecules is shown in Figure 4.3A. Upon setting the filter parameters (Z-crossing value = -1, σ = 1) contours were traced around the boundaries of the detected objects (Figure 4.3B). Two objects were selected as reference objects that were used by RAZA^{PS} to search the volume for objects with similar dimensions.

RAZA^{PS} next calculated the reduced structural information (length of major, minor, and middle axes, 3D surface area and 3D volume) for those two references and used these parameters to find similar objects within the tomogram (Figure 4.3B). All objects in the synthetic 'truth set' were detected (Figure 4.3C) based on major, middle and minor axes, 3D surface area and 3D volume at +/- 2% tolerance.

Figure 4.3D shows surface rendered views of the detected objects and their major (blue), middle (green) and minor (red) axes. These all have the same point of intersect, which represents the 'center' of the object. While there are different definitions of centers of objects for complex shapes (e.g. center of mass), this 'center' is useful for the extraction of objects for further classification and sub-volume averaging purposes.



	0	2	4	6	8	10
All	0	3	7	7	11	12
3D Axes	3	6	9	11	12	12
3D Surface Area	3	7	10	11	11	12
Minor Axis	3	9	9	11	12	12
Major Axis	3	9	12	12	12	12
3D Volume	3	11	12	12	12	12
Middle Axis	4	12	12	12	12	12

Figure 4.4 Analysis of parameter threshold levels on the total no of detected objects.
The truth set searched (Figure 4.3A) contained 12 particles.

Having established that RAZA^{PS} could reliably detect object contours in the truth set, we next tested the sensitivity of RAZA^{PS} to the adjustment of threshold values for each of the 5 search parameters (Figure 4.4). The thresholds were adjusted between +/- 1-10% of the reference values of the middle axis (orange), the 3D volume (green), the major axis (cerulean), the minor axis (light blue), the 3D surface area (purple), all 3D axes (maroon) and all parameters (blue). This graph clearly shows that different parameters respond differently to their percentage increase in their threshold values (from 1-10%). In this example, particle detection appeared most sensitive to the increase in the middle axis threshold with a 2% increase resulting in the detection of all 12 particles. Under the most stringent conditions in which all 5 parameters were controlled simultaneously all 12 particles shown in (Figure 4.3A) were detected at the 10% threshold. In comparison, a surface area threshold of 8% enabled the detection of all 12 objects.

Based on this analysis we conclude that despite the use of 12 identical truth objects in random orientations, RAZA^{PS} required approximately a +/- 2-10% threshold range around the 100% reference values to detect all objects accurately. One reason for this is that the approximation

methodologies used to generate the rotated objects can result in loss of information (Park, Leibon et al. 2009). Our results suggest a possible loss of visually un-noticeable information after rotation due to uniform noise.

Table 4.1 Structural details of GroEL detected at a 2% tolerance threshold.

We used 'Middle Axis' as search parameter. The structural parameters of two reference objects were averaged to calculate the % deviation. Deviation = ((value obtained from detected particle – value obtained from reference particle) / value obtained from reference particle)*100). Maj, Mid and Min represents major, middle and minor axes respectively.

	Axes Length (µm)						3D Surface Area (µm^2)		3D Volume (µm^3)	
Object Number	Maj	% Dev from Ref	Mid	% Dev from Ref	Min	% Dev from Ref	SA	% Dev from Ref	V	% Dev from Ref
1	97	1.44	82	11.90	72	9.52	34312	7.85	266760	1.80
2	94	1.69	87	6.53	76	4.49	36171	2.86	265887	2.12
3	95	0.65	84	9.76	81	1.79	38439	3.23	275658	1.48
4	95	0.65	90	3.31	76	4.49	38958	4.63	269640	0.74
5	97	1.44	86	7.61	76	4.49	37428	0.52	267405	1.56
6	95	0.65	85	8.68	76	4.49	36030	3.23	267622	1.48
7	95	0.65	88	5.46	75	5.75	37822	1.58	267201	1.63
8	95	0.65	84	9.76	81	1.79	38439	3.23	275658	1.48
9	97	1.44	79	15.13	77	3.24	35729	4.04	265839	2.14
10	96	0.40	86	7.61	78	1.98	39506	6.10	269245	0.88
11	98	2.49	89	4.38	76	4.49	41693	11.97	277999	2.34
12	98	2.49	89	4.38	76	4.49	39696	6.61	258963	4.67
Mean	**96**	**1.22**	**86**	**7.88**	**77**	**4.25**	**37852**	**4.65**	**268990**	**1.86**

Table 4.1 lists the respective mathematical fingerprints (length of major, middle and minor axes, surface area and volume) and their deviation from the parameters derived from the reference objects. By setting a 2% tolerance threshold, we found that on average, the detected length of major, middle and minor axes get respectively 1.22%, 7.88% and 4.25% deviation from the reference objects. Similarly the respective deviation of surface area and volume from reference objects were 4.65% and 1.86% respectively. Note that we used only the middle axis to search for similar objects. Because all objects are rotated versions of the

same GroEL molecule, ideally we would expect 0% deviation from reference object. The slight deviation observed may be due to the fact that the rotation algorithms use approximation criteria. These results suggest that RAZA^{PS} accurately detects all objects of GroEL volumes with a maximum threshold of 2%.

4.3.2 Detecting ribosome in a subcellular tomogram

Having established that RAZA^{PS} could accurately select all 12 GroEL molecules from the synthetic truth set, the next test was designed to establish whether RAZA^{PS} based *structural fingerprinting*, could be used to enable the selective extraction of relatively uniform classes of objects in cellular tomograms.

For this purpose, a chloroplast region of a high pressure frozen, freeze substituted plastic-embedded *C. reinhardtii* cell (Ali, Landsberg et al. 2012) was selected as it contained a multitude of black densities presumed from their size and subcellular location to be ribosomes.

Using RAZA^{PS} 10 suitable 'reference' ribosome-like molecules (Figure 4.5; z= -1 and σ=0.8) were identified and based on these, average values were calculated for their three axes, surface area and volume. Under the most stringent conditions tested (all three axes, surface area to volume set to +/-20% of the reference values) only 55 particles (Figure 4.5:A1-3) were obtained. Figure 4.5:B1-3 shows that when only the three axes parameters were set +/-20% of the reference values a larger number of particles (144) were detected as the search was less stringently controlled (i.e. no restriction on surface area or volume). When conditions were relaxed further by setting only the surface area to +/-20% of the threshold values 319 particles were obtained (Figure 4.5:CA1-3). Finally when only the major axis threshold was controlled and increased to 100%, all of the 4509 black objects within the tomogram were detected (Figure 4.5:DA1-3).

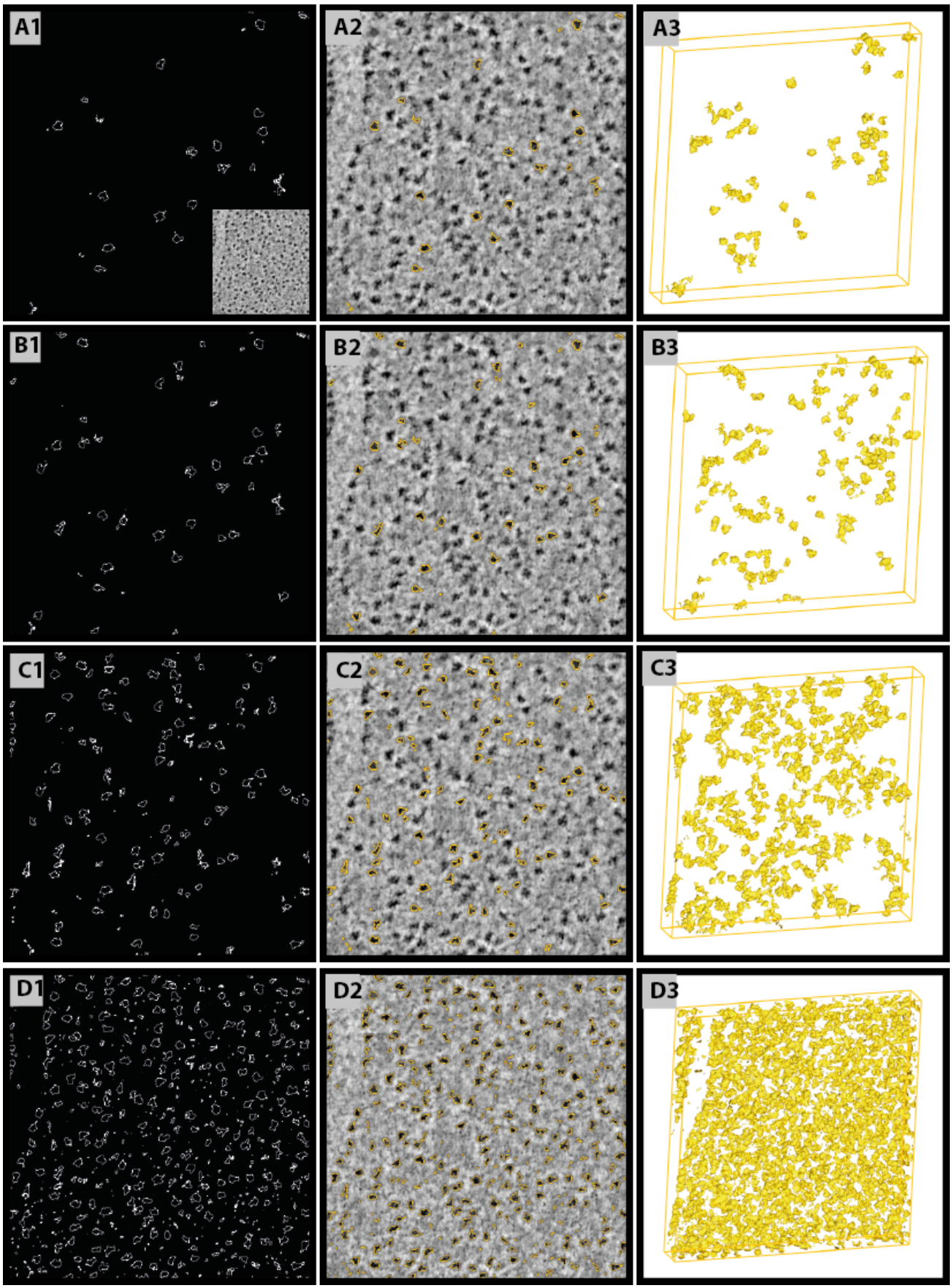
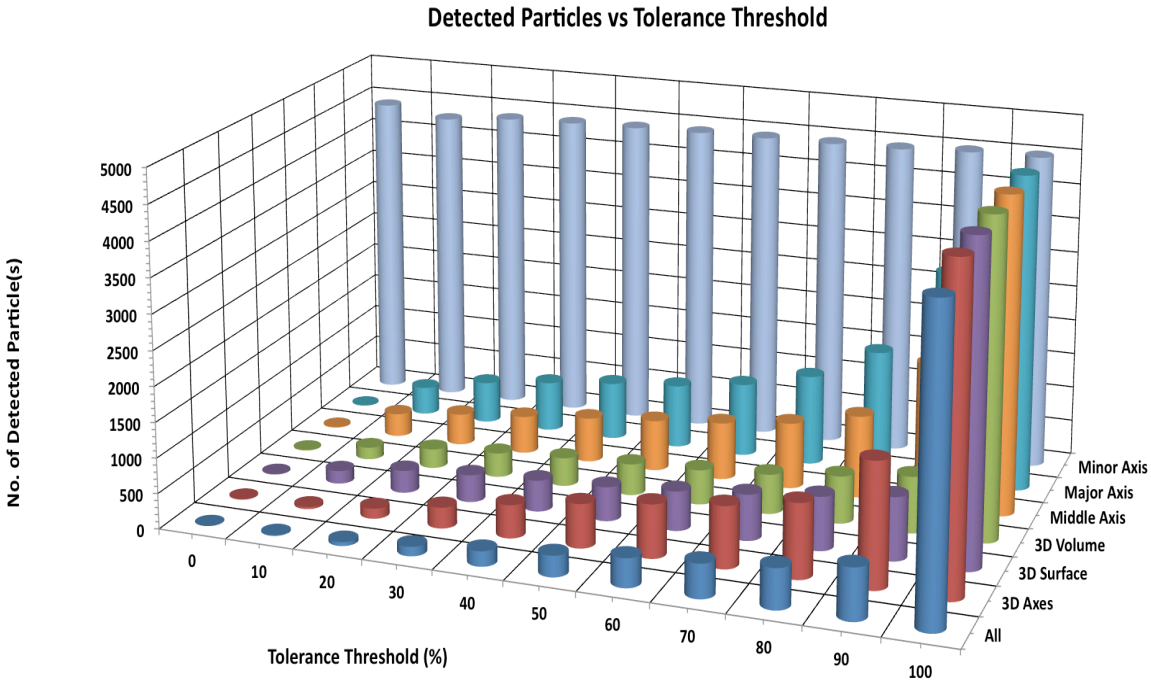


Figure 4.5 Detection of ribosome like macromolecular assemblies in cellular tomograms.

A1. Application of RAZA^{PS} at a 20% threshold setting for all three axes, Surface area and volume. A2 Enlarged inset from A1 with contours of detected objects detected. A3 3D surface rendered views of the detected objects. B1. Application of RAZA^{PS} at a 20% threshold setting for only the 3D axes. B2 Enlarged

inset from B1 with contours of detected objects. B3 3D surface rendered views of the detected objects. C1. Application of RAZA^{PS} at a 20% threshold setting for the surface area. C2 Enlarged inset from A1 with contours of detected objects.C3 3D surface rendered views of the detected objects. D1. Application of RAZAPS at a 100% threshold setting for the major axis. D2 Enlarged inset from A1 with contours of detected objects D3 3D surface rendered views of the detected objects.



	0	10	20	30	40	50	60	70	80	90	100
All	0	12	55	131	213	296	406	478	565	720	4288
3D Axes	9	31	144	296	468	620	752	867	1046	1744	4509
3D Surface Area	10	184	319	388	443	484	557	648	756	887	4508
3D Volume	9	171	275	345	404	446	493	562	670	798	4507
Middle Axis	12	332	451	541	640	728	819	941	1168	2028	4509
Major Axis	9	401	597	717	825	909	1051	1291	1761	3012	4509
Minor Axis	4412	4282	4367	4394	4409	4426	4433	4439	4450	4497	4509

Figure 4.6 Analysis of total no of detected ribosome-like objects.

The chart shows detection of ribosome-like objects at increasing levels of threshold tolerance based on major, middle and minor axes (separately) and together (3D axes), 3D surface area and 3D volume.

Figure 4.6 provides a comprehensive overview of the effect of adjusting individual and combinations of parameters between +/- 0 to +/- 100% of the threshold values. The thresholding of the minor axis in this example, appear to have little effect on the number of particles detected. However, in all other cases the numbers of particles detected increased with the threshold setting. Furthermore it should be noted that the adjustment of thresholds for the major and middle axes had a more pronounced effect (that is, detected more objects) than those obtained for the 3D volume and surface area of the objects. This may be explained by the fact that lengths are squarely and cubic proportional to surface are and volume respectively. Controlling 'all' parameters simultaneously

resulted in the most stringent control of particle selection. This suggests that initial search parameters to identify regularly shaped objects (e.g. ribosomes) might be achieved by setting the major axis to 100%. The stringency of the search could then be refined either through the reduction of the tolerance threshold or by the inclusion of other thresholded parameters (e.g. all). Middle and minor axes can have more object specific response. Although the conditions to identify each target object may require tuning, this experiment establishes clear proof of principle for the high throughput detection of regular objects as small as ribosomes.

4.3.3 Detecting mitochondria in tomograms

The next challenge was to establish whether the *structural fingerprinting* and *selective thresholding* capability of RAZA^{PS} could be used to automate the detection and segmentation of irregularly shaped objects such as mitochondria in cellular tomograms.

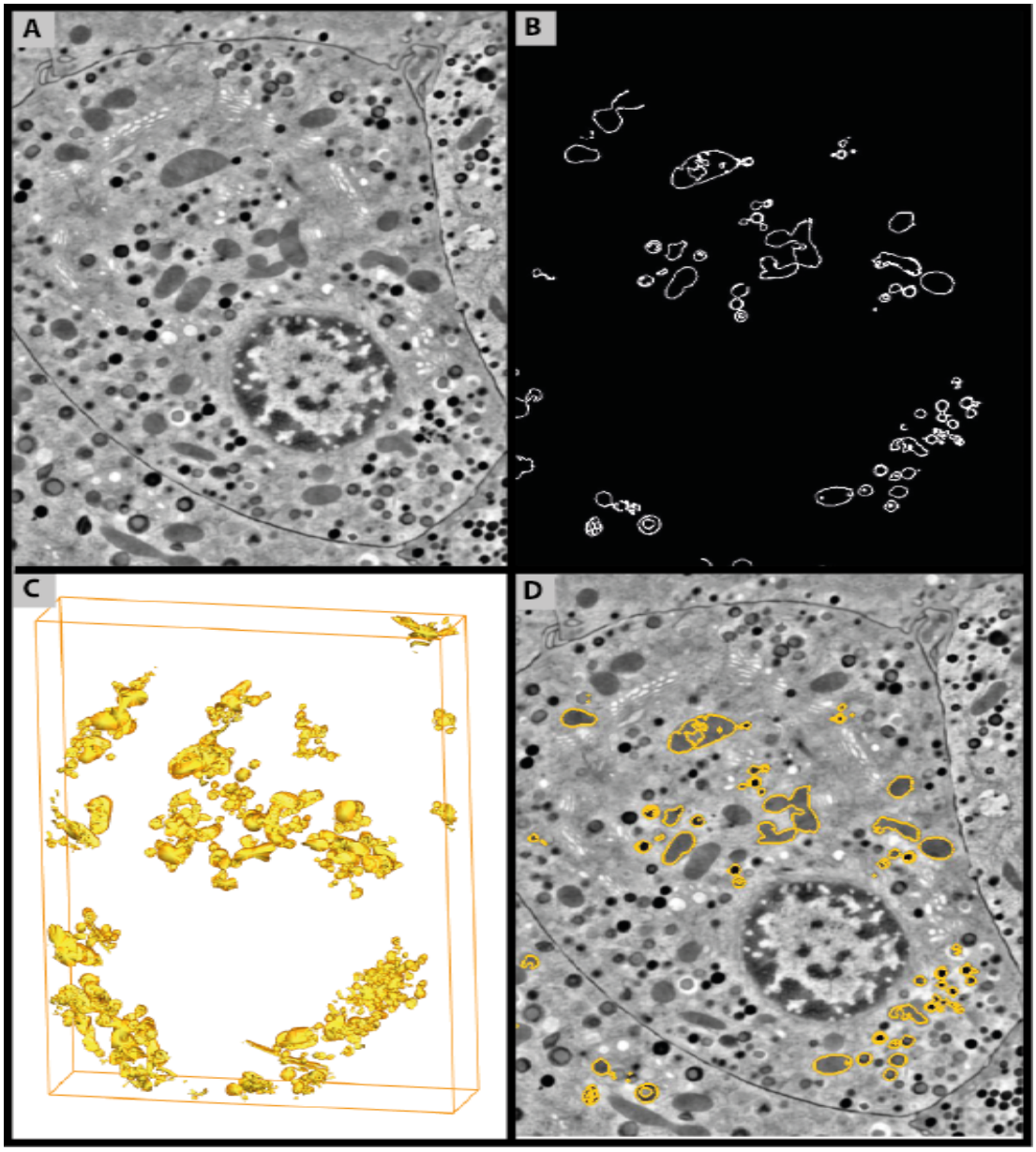


Figure 4.7 RAZA^{PS} based Mitochondria detection.

A) Tomographic section of a high-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell. B) Contours detected by RAZA^{PS} using Z-crossing value = 1170 and σ = 0.49 Arrowed contours correspond to reference objects. C) 3D surfaces rendered objects detected when the thresholds were set to 20% for all parameters (major, middle and minor axes, surface area and volume. D). Contours overlaid on the original gray scale tomogram.

An experienced analyst can identify mitochondria by their shape, size and appearance. This suggests that appropriate pattern recognition algorithms should also be able to do so. But while macromolecules such as ribosomes are quite regular in size, the size, shape and structure of mitochondria can vary significantly. Figure 4.7A shows a slice of a 3D cell tomogram containing different organelles, including mitochondria. The mitochondria of interest are in Figure 4.7A. When RAZA^{PS} was applied to this data (Figure 4.7B: Z-crossing value = -1170 and σ = 0.49) all traced objects detected were mitochondria, although not all mitochondria were detected. This was expected due to the structural variation of mitochondria. Five of these detected objects were used as reference objects to calculate average search parameters for the three axis lengths (major, middle and minor), surface area and volume. Based on these values, thresholds were adjusted to 20% for all five parameters. Taking this approach, 64 objects were detected (Figure 4.7C). The complexity of these objects is seen in Figure 4.7C, with the contours overlaid on the tomogram in Figure 4.7D.

These preliminary results show that while mitochondria exhibit significant variability in size, shape and structure, RAZA^{PS} is still capable of discriminately detecting a significant subset of mitochondria within a complex cellular volume. Note that RAZA^{PS} is robust in detection of mitochondria in tomograms not only because of its edge detection capabilities but also its capability of defining structural fingerprints. We also note that the variability in the size of mitochondrion objects requires a tolerance threshold of 20%. It is also clear from Figure 4.7, there are few mitochondrion segments that were not picked by RAZA^{PS}, a possible reason is interconnection of few mitochondrion sections due to technical limitations. RAZA^{PS} considers such sections as a single object, and calculates its structural parameters from that, and ignores them while searching the objects similar to reference volume. An indirect approach to detect such objects is by giving them in reference volume, however we believe that the structural parameters (major, minor, middle axes, volume and surface area) of these objects could be misleading. Another reason for not few detecting mitochondrion objects could be low signal to noise ratio (poor contrast) between close objects in the tomogram when applying RAZA.

4.3.4 RAZA^{PS} detection different objects with similar shapes

To determine the accuracy of RAZA^{PS} a synthetic truth set containing (20 copies of GroEL chaperonin complex (EMDB accession code 1081) in different orientations, 64 small sized spheres, 28 large-sized spheres) was constructed and contaminated with 60% Gaussian noise as discussed in

the Methods (section 4.2). The experimental aim was to evaluate, whether RAZA^{PS} is capable of accurately selecting the target object of interest, from a mixture objects including objects with similar object parameters. The object parameters of the modelled GroEL, small and large spheres are listed in Table 4.2. Note that the Major and Middle axes parameters of GroEL and the large spheres are within 9% of each other providing a stringent test for object identification on the basis of this parameter.

Table 4.2 Structural measurements of objects.

	GroEL	**Large Spheres**	**Small Spheres**	**Difference b/w GroEL and large spheres (%)**
Major axis (μm)	94	103	54	9
Middle axis (μm)	91	100	52	9
Minor axis (μm)	73	99	50	26
Surface area (μm^2)	37050	50320	12095	26
Volume (μm^3)	226199	471506	61453	52

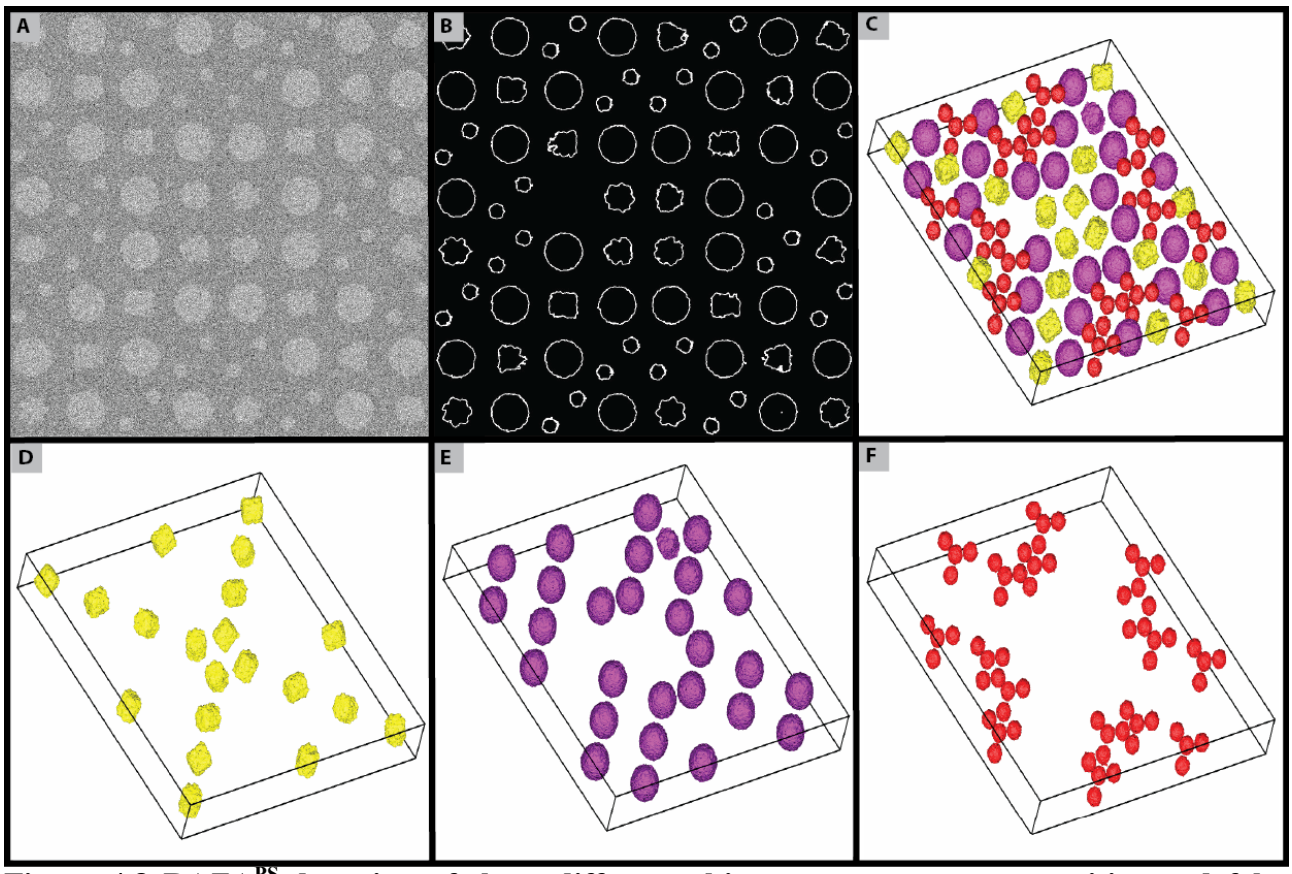


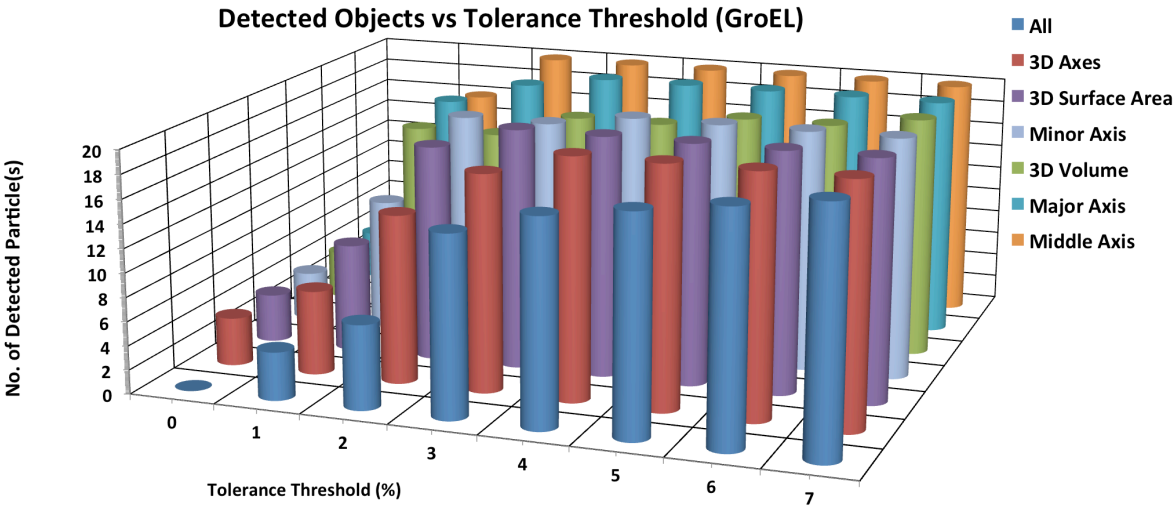
Figure 4.8 RAZA^{PS} detection of three different objects to measure true positive and false positive rates.

(A) Input 2D slice selected from 3D volume (containing 20 randomly oriented GroEL structures, 28 Spheres and 64 small spheres) contaminated with 60% Gaussian noise. (B) Application of RAZA^{PS}, which detected all objects detected at 36 % major axis tolerance setting. (C) Combined 3D surface view of all three object types (GroEL (yellow), large spheres (purple) and small spheres (red)). (D) Four of the 20 noise corrupted GroEL molecules were selected as reference objects to RAZA^{PS}. Based on their parameters, 100% detection of all 20 GroEL molecules was achieved using RAZA^{PS} settings of Z value = 10, and σ = 1, at 7% threshold using the most stringent structural fingerprinting settings (i.e. all three axes, surface area and volume parameters had to be within the threshold settings). (E) 3D surface view of 100% detection of large-spheres with RAZA^{PS} settings Z value = 10, and σ = 1 at a 6% threshold when the search parameter was set to *major axis*. Four of the 28 large spheres were selected as reference objects. (F) 3D surface view of 97% detection of small spheres with RAZA^{PS} settings Z value = 10, and σ = 1 at a 36% threshold setting with minor axis as search parameters. The minor axis is often more affected by noise explaining the high threshold setting. Four of the 64 small spheres were selected as reference objects to RAZA^{PS}. The failure to detect one small sphere at a 36% tolerance threshold suggests that noise contamination of the minor axis, may have a greater influence on the detection of smaller objects.

Upon setting the filter parameters (Z value = 10, σ =1) contours were traced around the boundaries of the detected objects (Figure 4.8B). Four copies of each object type were selected as reference particles and their object parameters (Table 4.1) were used by RAZA^{PS} to search the whole 3D volume for objects similar to reference particles. This approach identified 3 populations of particles which are colour coded in Figure 4.8C and which could subsequently be selectively classified (Figure 4.8 D-F).

Figure 4.9 provides a comprehensive overview of adjusting all parameters individually or in

combination between +/- 0 to +/- 100% of the reference values (result shown here up to 7%, by which point all the true positive were detected. Closer analysis of the GroEL class of particles (Figure 4.9) shows that different threshold settings (typically in the range of 2-7% of the reference value) were sufficient to achieve 100% *sensitivity* (i.e. all true positive detected) and *specificity* (i.e. no false positive detected).



	0	1	2	3	4	5	6	7
All	0	4	7	15	17	18	19	20
3D Axes	4	7	14	18	20	20	20	20
3D Surface Area	4	9	18	20	20	20	20	20
Minor Axis	4	11	19	19	20	20	20	20
3D Volume	4	16	16	18	18	19	19	20
Major Axis	4	17	19	20	20	20	20	20
Middle Axis	4	16	20	20	20	20	20	20

Figure 4.9 Analysis of parameter threshold levels on the total no of detected objects for GroEL out of 20.

The Receiver operating characteristic (ROC) represents each point as a sensitivity (true positive)/specificity (false negative) pair corresponding to a particular decision threshold (tolerance threshold) and Area under ROC (AUROC) which defines how well a parameter can distinguish between two classification groups such as GroELs and non GroELs respectively (Figure 4.10). These show that when the false positive rate and true positive rate is perfectly balanced, the AUROC equals 1 which is achieved in the GroEL test. Here the x-axis represents the false positive level, whereas the y-axis shows the true positive level. As *sensitivity* is defined as the true positive object detection level, in this case detection of 20 GroELs gives a *sensitivity* of 100%, whereas *specificity* is the detection level of true negatives. In this case there were 112 objects in total, of which 92 were not GroELs and so 'true negatives'. Perfect detection of all true positives and exclusion of all true negatives yields a sensitivity=100%: 1-specificity=100%).

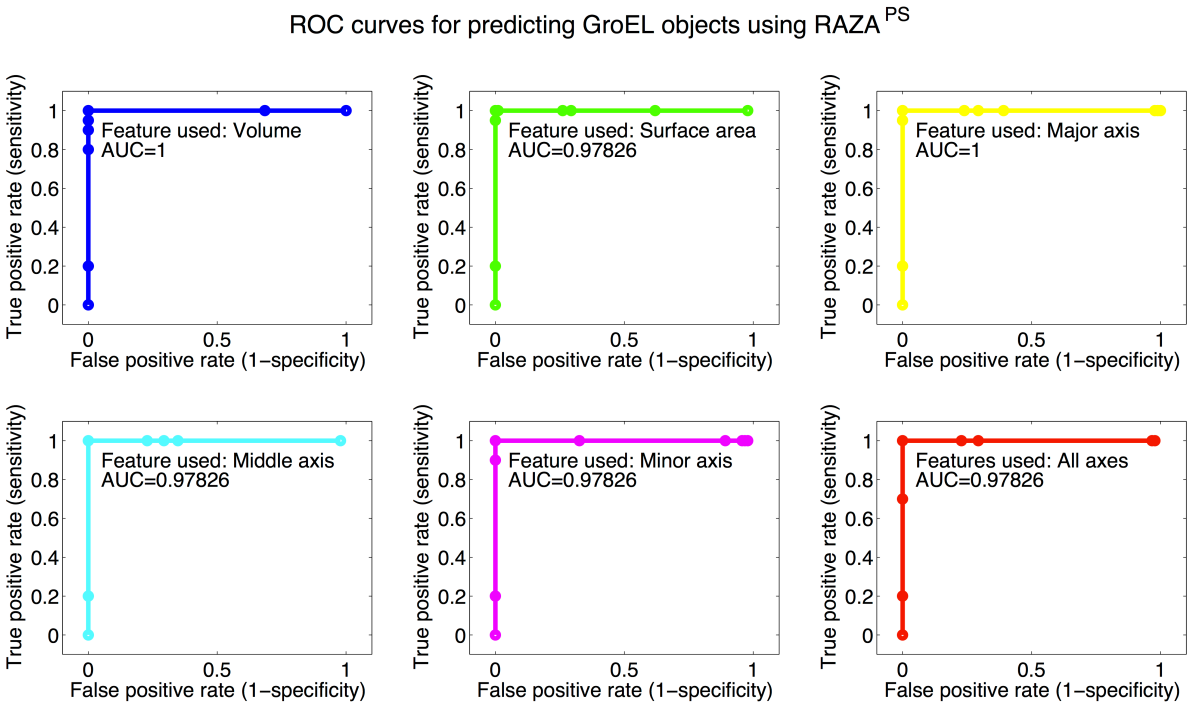
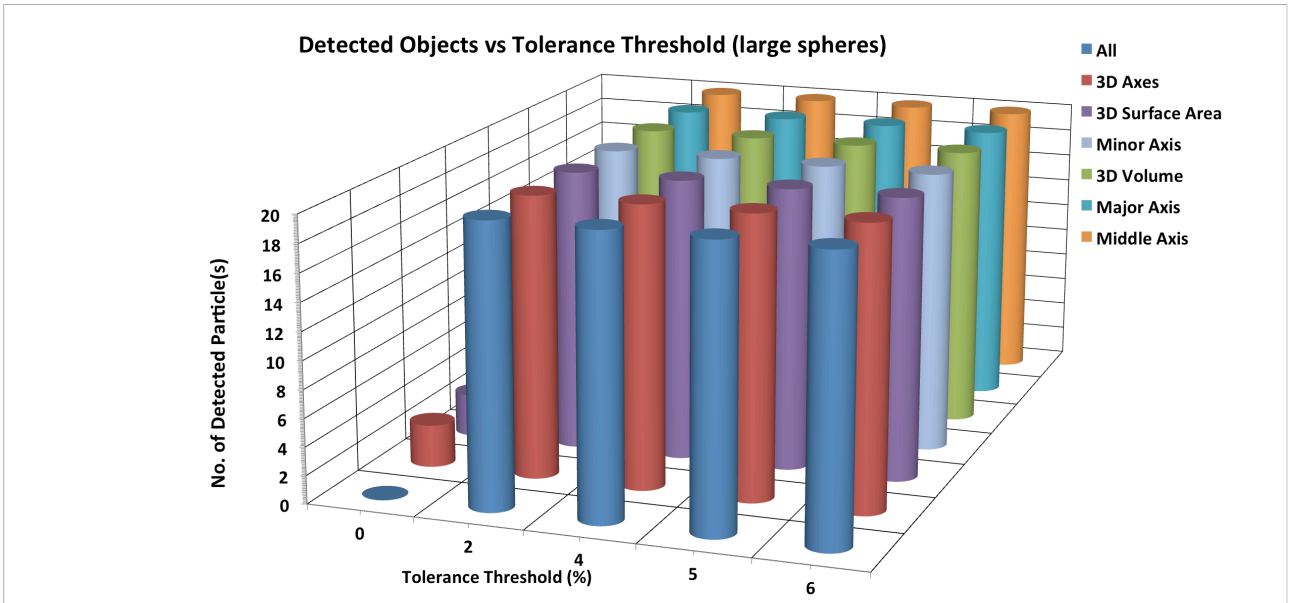


Figure 4.10 Receiver operator characteristic (ROC) curves for GroEL.

The order of optimise parameter settings in terms of the threshold limit required for 100% detection the following settings could be used: Major axis (3% threshold), Middle axis (2%), Minor axis (4%), Surface area (4%) and 3D axes (4%). The search parameters are conditions in which all 5 parameters were controlled simultaneously. The search parameters are conditions in which all 5 parameters were controlled simultaneously. The particle detection appeared most sensitive to the increase in the middle axis threshold, within a 2% increase resulting in the detection of all 20 GroELs. By identifying objects using ‘All” geometric parameters (most stringent case), a tolerance threshold of 7% proved to be sufficient to detect all GroEL particles with maximum AUROC=0.98.



	0	2	4	5	6
All	0	22	26	27	27
3D Axes	3	27	27	27	27
3D Surface Area	3	27	27	27	27
Minor Axis	3	27	27	27	27
3D Volume	3	27	27	27	27
Major Axis	3	27	27	27	28
Middle Axis	4	27	27	27	27

Figure 4.11 Analysis of parameter threshold levels on the total no of detected objects for large spheres (out of 28).

In the second test, we provided four large spheres as reference objects to be searched in the same test volume containing 112 objects in total containing 28 large spheres, whereas 84 were 'true negatives'. Under the most stringent conditions tested (all three axes, surface area to volume set to +/-50% of the reference values) all 28 spheres were obtained (Figure 4.11). In this figure details provided are up to 6% threshold only, as beyond this, false positives were detected. In each search case having '*all three axes combined*', '*3D Surface area*', '*Major axis*' and '*Minor axis*' as search parameters AUROC could achieve a 'perfect 1'. The *Major* and *Middle* axes proved to be less robust (more sensitive to noise) requiring detection thresholds of 6% and 7% respectively. Similar to the GroEL test, the parameter requiring the highest threshold was 3D volume (50%), whereas *Minor axis*, *Surface area* and *3D Axes* achieved 100% detection level at 20% threshold. Perfect detection of all true positives and exclusion of all true negatives yields a sensitivity=100%: 1-specificity=100%) resulting an AUROC equals 1 (Figure 4.12)

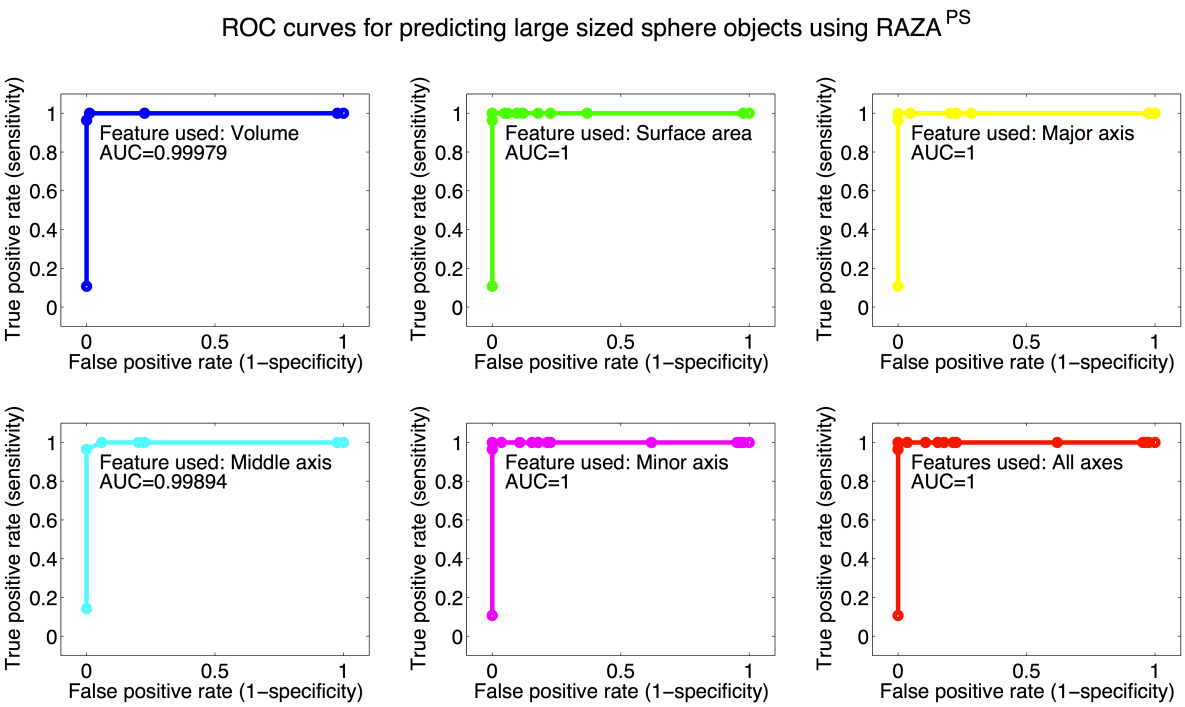
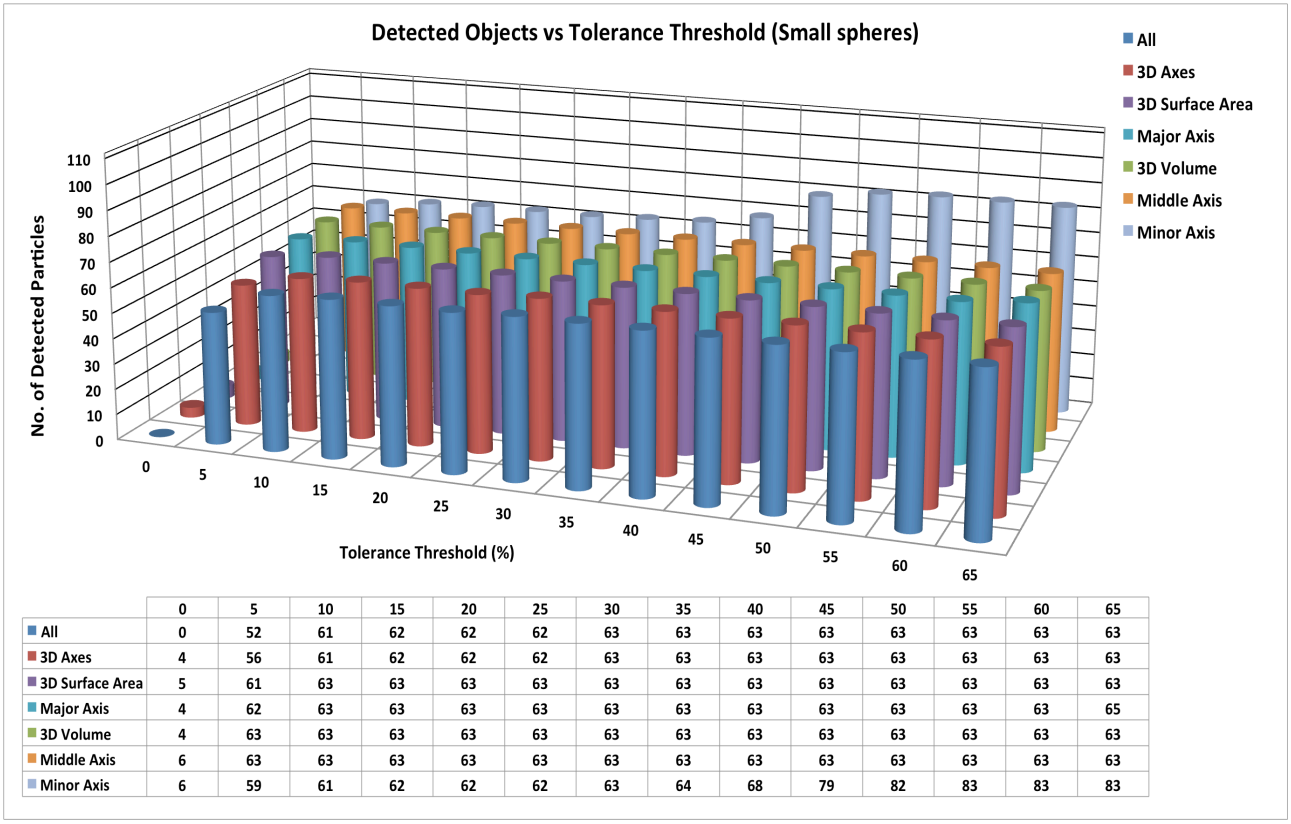


Figure 4.12 Receiver operator characteristic (ROC) curves for large spheres.

Similar to previous two tests, last test was conducted to detect the 64 small spheres from the 112 objects (leaving 48 as 'true negatives'). Using RAZA^{PS} noise corrupted small sphere references were provided to software to identify objects having same *fingerprints*.

The detection of the small spheres at increasing levels of threshold tolerance (within +/- 100% of reference values) was tested based on major, middle and minor axes (separately) and together (3D axes), 3D surface area and 3D volume. Figure 4.13 provides details about number of small spheres detected at each threshold. Summarising the details, again the most stringent test was considered to be the combined control of surface area, volume, major, middle and minor axes. The highest threshold (95%) was required when search parameters were set 3D Volume and 3D Surface area. Search based on Minor axis was shown to be more robust and detected all objects at a ~40%. When *All 3 axes*, *Major axis* and *minor axis* were controlled and conditions were relaxed further by setting threshold up to +/- 70% of the reference values, all the small spheres were detected at 60, 65 and 70% respectively. Again, perfect detection of all true positives and exclusion of all true negatives yields a sensitivity=100%: 1-specificity=100%) resulting an AUROC equals 1 (Figure 4.14).



	0	5	10	15	20	25	30	35	40	45	50	55	60	65
All	0	52	61	62	62	62	63	63	63	63	63	63	63	63
3D Axes	4	56	61	62	62	62	63	63	63	63	63	63	63	63
3D Surface Area	5	61	63	63	63	63	63	63	63	63	63	63	63	63
Major Axis	4	62	63	63	63	63	63	63	63	63	63	63	63	65
3D Volume	4	63	63	63	63	63	63	63	63	63	63	63	63	63
Middle Axis	6	63	63	63	63	63	63	63	63	63	63	63	63	63
Minor Axis	6	59	61	62	62	62	63	64	68	79	82	83	83	83

Figure 4.13 Analysis of parameter threshold levels on the total no of detected objects for small spheres (out of 64).

The chart shows detection of ribosome-like objects at increasing levels of threshold tolerance based on major, middle and minor axes (separately) and together (3D axes), 3D surface area and 3D volume.

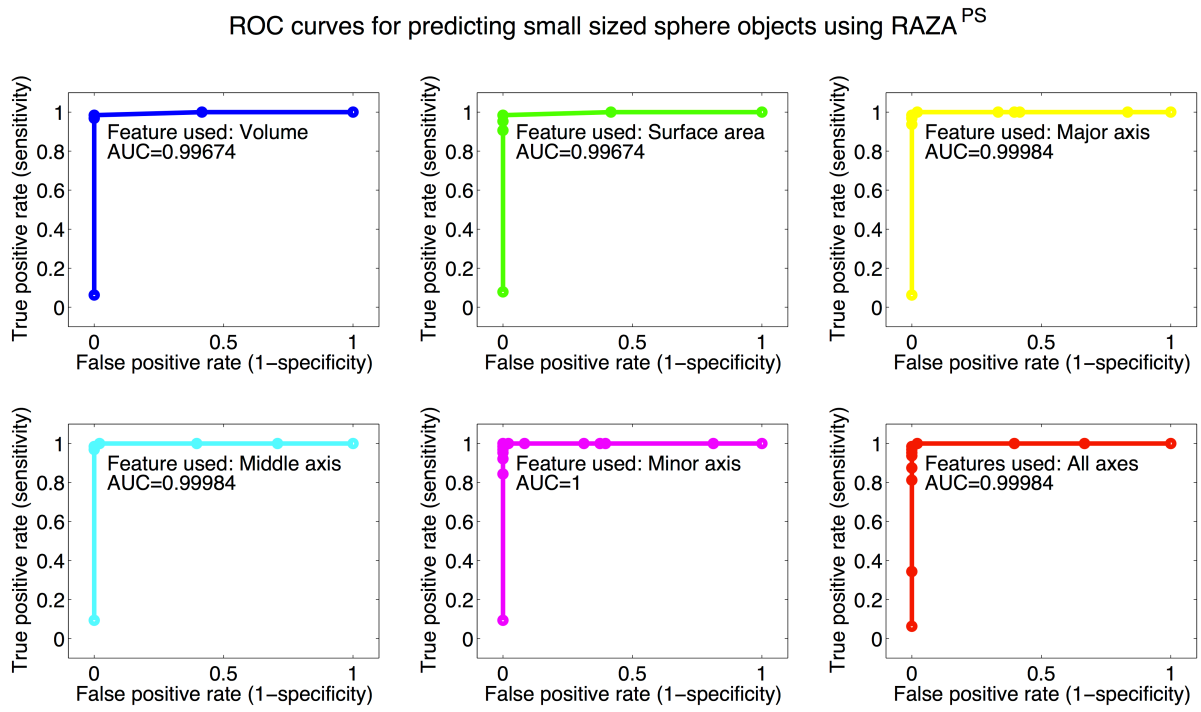


Figure 4.14 Receiver operator characteristic (ROC) curves for small Spheres.

4.3.5 Standard adaptation of 2D LoG Z-crossing based particle selection to 3D is not accurate as compared to RAZAps.

In this Chapter, the novel RAZA particle selector RAZA^{PS} component of RAZA was developed and tested. This chapter aims to provide structural biologists with a rapid, accurate and robust automated approach not only for high throughput detection of edges that define organelles, macromolecular assemblies and membrane proteins in cellular tomograms, but also the detailed structural measurements e.g. length, width, height, 3D surface area and 3D volume. We initially tested RAZA^{PS} on three different 3D volumes, which represent a broad range of typical microscopy datasets.

To our knowledge the structural fingerprint based particle selection and extraction has not been previously reported in literature. The novelty added by RAZA^{PS} is not achieved just by a simple adaptation of 2D to 3D.

A standard and simple adaptation of 2D LoG z-crossing based particle selection is to run 2D algorithm on each slice to determine length major and minor axis. The length of these axes can either be averaged or selectively used based on user's experience. These values can be used to search whole tomogram to find particle of interest.

To show this, we used 2D version of LoG based z-crossing algorithm to determine major axis length and minor axis length of particles of interest. We then used values obtained by this 2D method and searched in 3D volume for particle selection. The results are shown in Figure 4.15.

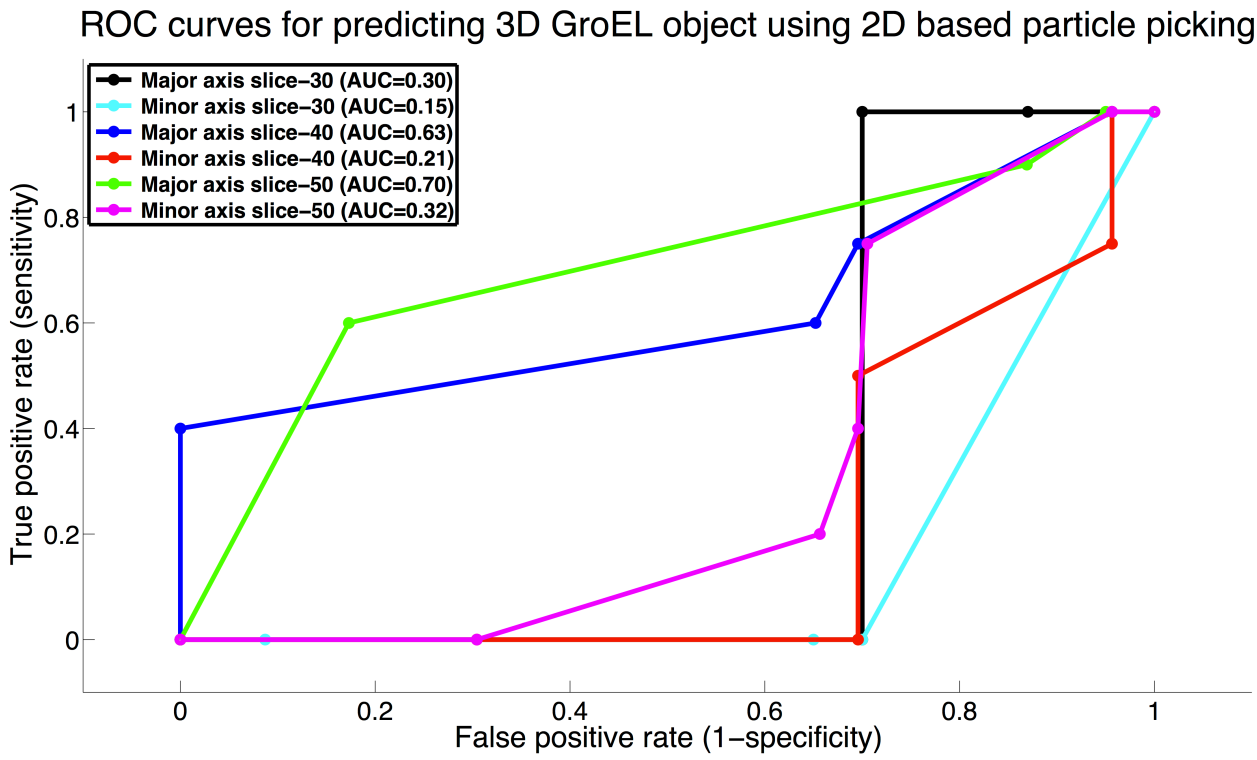


Figure 4.15 ROC curves representing standard adaptation of 2D algorithm in 3D Particle selection is based on slice 30, 40 and 50 for major axis length and minor axis length.

In three different tests, I used three different slices (slice 30, 40 and 50) and ran LoG based z-crossing algorithm to determine lengths of major and minor axes of reference particles within each 2D slice. When using 'slice 30' major and minor axes were given as search parameters, AUROC of 0.51 and 0.11 (Figure 4.15). Similarly in the second test, when 'slice 40' was used to determine geometric parameters of reference particles the AUROC of 0.63 and 0.21 was determined by using major and minor axis respectively as search parameters. In the final test, 'slice 50' was used to determine the geometric parameters of the reference particles. AUROC of 0.70 and 0.32 was achieved by using the major and minor axes respectively as search parameters.
As expected the above results show that a simple adaptation of 2D to pseudo 3D is not efficient enough for particle selection in 3D as the true maximum, middle and minor axes lengths cannot be determined and so the novel 3D RAZA^{PS} provides a clear advantage.

4.3.5 Technical limitations of RAZA^{PS}

The performance of RAZA has been extensively tested on different datasets (Figures 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 4.3, 4.4, 4.5, 4.7, 4.8) and in all of these cases it performed well. Through these analyses it was shown to be capable of accurately segmenting membranes, organelles, membrane proteins and macromolecular assemblies in a range of conditions including cryo-tomograms and resin embedded, freeze substituted samples collected in a range of laboratories. In addition it successfully identified all true positives and excluded all true negatives in a truth set (corrupted with 60% Gaussian noise) populated with three different populations of GroELs, small spheres and large spheres (Figure 4.8).

Low contrast images are clearly more challenging that high contrast images. However as is shown in Figure 3.8 particles which contain macromolecules similar in size to ribosomes can be detected in cryo-tomograms and this is only expected to improve with imaging technology. Consequently current cryo-tomograms provide a sensible 'worst case scenario'.

Technically edge detection is limited by tomogram signal to noise ratio and the setting of sigma and Z parameters. Small sigma values reduce smoothing and are typically used to detect small objects such as macromolecules and membrane proteins, where maximum accuracy is required. For the detection of large objects e.g., cell or organelle membranes larger sigma values can be used to smooth out noise. By measuring the absolute pixel intensity values across an edge to be segmented the first order derivative data can be obtained and used to calculate second order derivatives (or z-values), which allows careful parameter tuning. In addition to that due to the fundamentally important advances at the level of imaging, data capture and image processing (discussed in section

1.4.3) it is possible to achieve ~4nm resolution in vitreous ice samples (Engel, Schaffer et al. 2015). So, given these advances in technologies and the increased contrast that they deliver it is anticipated that RAZAPS mediated segmentation should improve further in future.

4.4 Conclusion

The advanced *Rapid, Automated Z-crossing algorithm* for *Particle Selection* ($RAZA^{PS}$) has been shown to be a high-throughput Laplacian-of-Gaussian edge-detector which can rapidly output discrete mathematically defined object contours. We have shown that for randomly oriented but otherwise identical molecules embedded in a volume corrupted with synthetic noise, the longest axis (*major axis*), a second longest axis orthogonal to this (*middle axis*) and the third orthogonal axis to these (*minor axis*) as well as a calculated surface area and volume of individual objects could be measured to within 10% of reference values. These 5 parameters were successfully used as object identifiers (*structural fingerprints*). Calculating these at the object level has the major advantage that it overcomes the problem of the 'random' orientation of subcellular structures (e.g. of ribosomes in cells).

Ribosomes are essential for protein synthesis. Detecting the ribosomal structure can help us to understand its structure-function relation for translating mRNA to protein. The densities of ribosomes are good in contrast but over-crowded in the tomogram. $RAZA^{PS}$ can be configured to segment all detected objects within a tomogram, or a selected subpopulation of objects (e.g. mitochondria or ribosomes). In Chapter 4, it was demonstrated that the specificity of object selection could be fine-tuned based on the Gaussian radius (δ), the z-crossing value and the *structural fingerprint* parameters. Biological diversity was effectively accounted for by tuning individual threshold settings for each of the 5 structural fingerprint parameters (e.g. 70-130% of the major, middle and minor axes, surface area and/or volume). The number of particles selected in the examples tested was shown to be most sensitive to the threshold setting for single parameters (e.g. length). Using all 5 *structural fingerprint* parameters provides maximum selectivity as collectively they provide an approximation of the structure of interest. For example, when searching for ribosomes, objects corresponding to organelles can largely be excluded through the use of any of the above parameters as organelles are generally larger in terms of their 3 axes, surface area and volume. In contrast ribosomes and proteasomes for example are much more similar in their overall dimensions, and so all five parameters may be needed to distinguish between them. Furthermore, when segmenting specific subcellular structures (e.g. microtubules), their selective detection can be assisted by searching for objects with a major axis (corresponding to tubule length) and two smaller axes (corresponding to tubular cross section).

RAZA^{PS} also calculates object centers, the location of which is defined as the intersection of the 3 defined axes. While there are alternative definitions of object centers (e.g. mass center) the current definition is suitable for rapid high-throughput sub-volume extraction as it minimizes additional center calculation times and because precise centering is not essential for downstream sub-volume averaging.

RAZA^{PS} also offers significant potential for the construction of 'atomic resolution' 3D models of cells as the molecular contours of their constituent objects and their *structural fingerprints* provides a framework for semi-automated docking of atomic protein structures into cellular tomograms. With the current 'resolution revolution' in cryo-EM facilitated by advances in direct electron detectors (Kuhlbrandt 2014), the number of atomic resolution membrane protein and macromolecular structures being solved is rapidly increasing. These atomic structures can also be used to define the major, middle and minor axes, surface area and volumes of target molecules, therefore providing validated search parameters for RAZA^{PS} to locate corresponding objects in cells. The high detective quantum efficiency of direct electron detectors (McMullan, Faruqi et al. 2014), particularly in single electron counting mode may well open up the opportunity not only to detect organelles and macromolecular assemblies but also membrane proteins in membranes, as there are already examples in the literature that this is achievable in isolated membranes (Kouril, Dekker et al. 2012). Consequently RAZA^{PS} provides a bridge to refining our understanding of the 3D molecular architecture of cells including the *in situ* organization of organelles, macromolecular assemblies and membrane proteins, which is now at the forefront of structural biology. Furthermore as RAZA^{PS} is designed to conduct automated Laplacian of Gaussian (LoG) edge detection it is largely scale and technology agnostic and so could likely have significant applications in medical imaging (e.g. MRI and CAT) and beyond.

CHAPTER 5

Conclusions and Future work

5.1 Research aims revisited

In conclusion below I have summarized by achievements in terms of my stated aims

Aim 1: To evaluate (review) available imaging technologies. To identify how available methods can best be used to perform noise reduction and image segmentation in 3D to help define structure of macromolecular complexes.

The aim of Chapter 1 was to summarize the importance of microalgae and their photosynthetic machinery in terms of solar energy capture, as well as the importance of structural biology to enable structure guided design of next generation microalgae and potentially, artificial photosynthetic systems. This was followed by a summary of advances in structural biology techniques and in particular electron microscopy. Some of the key factors affecting final image resolution including the electron source (e.g. Tungsten filament, Lanthanum hexaboride (LAB_6) filament and field emission gun (FEG)) through to direct electron cameras were covered, as were measures of image quality assessment.

Noise reduction methods are vital for removing the unwanted information in an image prior to segmentation. I therefore surveyed the main classes of noise reduction algorithms and specific examples of these. They include low pass fast Fourier transform, filter based noise reduction filters, wavelet transformation, non-linear anisotropic diffusion (NAD), iterative median, bilateral and discriminative bilateral (DBL) filtering. Most of these filters have been implemented in 2D. I also evaluated these noise reduction methods on the basis of their operating parameters, computational and space complexity. For this thesis, I selected Gaussian followed by median filtering for pre-processing of electron tomograms. In some cases I also used NAD filtering. The rationale behind

the use of these filters is that they do not add extra data or remove significant parts of the image to facilitate pre-filtering prior to edge detection and segmentation.

Segmentation in 3D is routinely performed manually, but this is very time consuming and requires a level of biological expertise. For example, the 3.1 x 3.2 x 1.2 μm^3 volume presented in (Marsh, Mastronarde et al. 2001) required around 9-12 months of manual segmentation using IMOD (Hohn, Tang et al. 2007). Semi-automated segmentation is usually less laborious than pure manual segmentation, and can easily segment larger compartments, however the segmentation of small objects requires improved accuracy and ideally a high degree of automation.

A number of algorithms have been employed for semi-automated segmentation including linear interpolation, thresholding image intensity to binarize the volume. Snake and watershed algorithms are also routinely used. In Chapter 1, I therefore provided an example of using semi-automated methods to segment an algal cell. I showed that larger objects such as the chloroplast and nucleus could easily be semi-automatically segmented. However, the semi-automatic segmentation of the eye spot, mitochondria, starch granules, Golgi and ribosomes was still very time consuming.

Edge detection methods were also reviewed. These techniques include the use of Sobel operations, the Laplacian of Gaussian (LoG), the Canny edge detector, the bilateral edge filter (BLE) and Marr-Hildreth filtering. For electron tomographic data, most of the available edge detectors operate in 2D (with the exception of the recursive edge detector that works in pseudo 3D fashion and Canny edge detector which is only currently implemented in 3D). The Canny edge detector (Canny 1986) and bilateral edge filters (Pantelic, Ericksson et al. 2007) appear to be significantly more accurate than other techniques (Marr and Hildreth 1980, Woolford, Hankamer et al. 2007). While the Canny and bilateral edge filter offer similar levels of performance in terms of tomogram segmentation, the bilateral edge filter only requires the setting of one adjustable parameter as opposed to three for the Canny edge detector (Pantelic, Ericksson et al. 2007). For electron tomographic data, I emphasized that there is a need to develop automated, high-accuracy 3D edge detectors that have the capability to suppress noise and maintain true edges. Therefore one of the major goals of this thesis was to develop such methods.

Aim 2: “To develop, test and evaluate a 3D noise reduction and edge detection filter that is capable of resolving the contours of individual macromolecular assemblies in the cell”.

In Chapter 2 and 3, we developed two novel edge detection methods; the 3D bilateral edge detector (3D BLE) and RAZA edge detector, respectively. The 3D BLE (Ali, Landsberg et al. 2012) is capable of suppressing noise and detecting objects of interest. The accuracy of the 3D BLE filter

was evaluated using a range of synthetic and real biological data sets. On synthetic 'truth' reference volumes containing hollow cylindrical, spherical, triangular and rectangular objects contaminated with different combinations of Gaussian, impulse and real tomographic noise, the 3D BLE outperformed the recursive edge detector (Monga, Deriche et al. 1991) and yielded a slightly better performance than Canny edge detector (Canny 1986).

The 3D BLE filter was also tested using a simulated cryo-tomogram populated with 100 uniquely oriented copies of the GroEL chaperonin complex. The 3D BLE was again shown to perform better than the 3D recursive edge detector and the Canny edge detector using this data set. Similarly, when testing real biological tomographic data sets, 3D BLE outperformed both the Canny and recursive edge detectors. For example, on an 800 × 800 × 100 voxel region of a dual tilt tomographic reconstruction recorded from a resin-embedded, sectioned and post-stained *C. reinhardtii* cell, the 3D BLE showed a greater ability to suppress noise, contour subcellular volumes and detect ribosome like macromolecules as small as ~25 nm in diameter. This confirms that the 3D BLE filter is capable of detecting true edges in a complex, crowded, noisy 3D environment. Although, I concluded that the 3D BLE is a better edge detector for automated electron tomogram segmentation than other methods currently available (Canny 1986, Monga, Deriche et al. 1991) the discontinuity in the detected edges is a disadvantage. These discontinuities are due to the fact that the 3D BLE filter marks various pixels as being part of an edge, but does not connect these pixels into a single mathematical object. Specifically in a tomographic data set corrupted with high levels of noise, the 3D BLE filter considers some of the true edge voxels to be noise, which results in discontinuous edges. As a result the objects defined by these edges cannot be reduced to a single mathematical object; specifically, the estimation of geometric properties such as length of orthogonal axes, surface areas, volumes and centers of objects cannot be accomplished if the object is not enclosed by a single, continuous contour. Although methods like spline interpolation can be used to connect contours, but these methods could generate spurious edges in a noisy and crowded environment of macromolecules. Additionally such methods are computationally expensive.

In Chapter 3, we developed the rapid, automated 3D Z-crossings algorithm (RAZA). RAZA combines the denoising operation inherent to Gaussian filtering with an automated edge detection capability of the Laplacian function. Through the process of LoG filtering followed by applying an arbitrary Z-crossings threshold, the algorithm facilitates not only the segmentation of objects, but also allows objects to be generalised as discrete mathematical objects as it yields continuous edges. This enables objects such as organelles, membranes, macromolecular assemblies and membrane proteins, which differ greatly in size and shape, to be detected. First RAZA was tested on a 300-400 nm section of high-pressure frozen, freeze-substituted and plastic-embedded mouse islet pancreatic cells that contain a range of subcellular compartments with different pixel intensity levels. By

tuning the values of Z-crossings and sigma, RAZA was able to selectively detect Golgi, mitochondria and mature insulin granules. Although both edge detectors, the 3D BLE and RAZA, could resolve the contours of macromolecular assemblies, RAZA was shown to have the added advantage of generating continuous 3D contours each of which effectively defines a discrete geometric volume with quantifiable properties (e.g. length, width, height, 3D volume and surface area) which can theoretically be used both for object detection (e.g. the selection of macromolecular assemblies in a particular size range) and potentially, automated cellular annotation and molecular docking.

Aim 3: "To compare manual segmentation with semi-automated segmentation facilitated by edge detection methods"

To compare manual segmentation with the performance of semi-automated segmentation by both 3D BLE and RAZA, a tomographic reconstruction of an insulin-secreting pancreatic beta cell was used as a test volume. In Chapter 2 and 3, it was shown that both are able to reproduce the structural elements identified by manual segmentation. The time and space complexity of the 3D BLE algorithm was also assessed, leading to the conclusion that the 3D BLE filter was ~4000 x faster than manual segmentation. In Chapter 3, I used the structural features identified by manual segmentation as benchmarks to evaluate the accuracy with which the same objects were detected by RAZA. I found that the surface area and volume of objects detected by RAZA was approximately the same for Golgi vesicle and mitochondria (less than 2 % deviation from reference structures). However for endolysosomal compartments the surface area and volume deviated up to 34% from the reference structures. A possible reason for the high difference between manual and RAZA mediated endolysosomal segmentation is that RAZA delineated the contours in a more detailed way and also included additional objects (true positives) missed by manual segmentation, which increases the total surface area and volume detected. RAZA was comparable to the 3D BLE filter in terms of time, almost 4000 x faster than manual segmentation.

Aim 4: "Detection of membrane proteins within tomographic data using edge detection methods "

The above-mentioned tests were based on conventional, negatively stained, resin-embedded electron tomographic datasets. Cryo-electron tomograms reveal the structure of specimens in their near native environment (i.e. they are neither stained nor fixed in any way). In practice, the resolution achieved by cryo tomography is not yet high enough to yield atomic resolution structure,

although with recent advances this may become possible (Kuhlbrandt 2014). Consequently, the detection of populations of individual membrane proteins in cryo-tomograms of membranes using the edge detection methods developed is an exciting possibility. In Chapter 3, I investigated the ability of RAZA to segment membranes and membrane-embedded proteins under low dose cryo-imaging conditions. The results show the capability of RAZA to conduct semi-automated segmentation of both membranes and large extracellular domains of membrane-embedded proteins. I further validated these findings by comparing the segmentation achieved by RAZA with that obtained through manual segmentation (Davies, Strauss et al. 2011). Using a 3D tomographic reconstruction of a 2D crystal of the V_oV_1 rotary ATPase isolated from negatively stained *Thermus thermophilus* membranes (Gerle, Tani et al. 2006), I showed that RAZA accurately detected both the extracellular and transmembrane regions of the V-ATPase in the reconstituted 2D crystal.

Aim 5: "To use edge detection methods to develop 3D particle-picking tools for the quantification of macromolecular assemblies and subcellular compartments by defining *structural fingerprints* using geometric properties of segmented objects."

One of the advantages of accurate edge detection methods that yield continuous contours is that it enables objects to be defined mathematically. Objects can be defined in terms of their longest axis (major axis), the next longest axis orthogonal to the major axis (middle axis) and the third orthogonal axis to both major and middle axes (minor axis), surface area and volume for each segmented object. These five parameters provide object specific *structural fingerprints,* which can be used for object identification, selective segmentation, classification and/or extraction.

In Chapter 4, I extended the RAZA edge detection method described in Chapter 3 by developing RAZA^{PS} (*Rapid, Automated Z-crossing algorithm for Particle Selection*). To achieve this, I used a connected component labelling approach to define and label each segmented object separately. I then calculated the lengths of major, middle and minor axes of each object in addition to its volume and surface area and set user-defined tolerance thresholds for each to allow some variability during the detection of objects of interest. RAZA^{PS} was tested on a simulated 3D volume containing 12, uniquely oriented GroEL particles. By providing two of these objects as references, RAZA^{PS} was able to detect all 12 objects of interest at a tolerance of 2%.

In order to test the particle selection functionality of RAZA^{PS} to detect ribosomes within a *C. reinhardtii* cell using *structural fingerprints*, a region of a high pressure frozen, freeze substituted plastic-embedded *C. reinhardtii* chloroplast containing a large numbers of ribosome like particles was analysed. Ten reference objects were used to allow RAZA^{PS} to define the 'structural fingerprints' to be used as search parameters. As these ribosome-like densities were very crowded

in the tomogram, in initial trials two adjacent densities were on occasion labelled as a single object. To overcome this, the tolerance thresholds were set to 100% of the reference length (i.e. a 0% tolerance), surface area and volume values. While this prevented the problem, the 100% setting resulted in RAZA^{PS} being overly stringent in terms of its selection process. By relaxing these thresholds, a total of ~ 4500 ribosome like molecules were rapidly contoured, identified and extracted to assist with their quantification such as their shape, size and abundance.

To provide crucial quantitative insights into the changes in the distribution, size and number of mitochondria, I used a tomogram from a high-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell isolated form Islets of Langerhans (Noske, Costin et al. 2008). Using tolerance rate of +/- 20% of the reference values based on five reference mitochondria, RAZA^{PS} segmented 64 mitochondria. I showed that mitochondria could be classified into different groups based on their 'structural fingerprints' i.e. length, height, width, surface area and volume. Furthermore the data enables subsequent principle component analysis to further this classification process. It was noted that a few mitochondria were not selected by RAZA^{PS} due to technical limitations of low SNR tomographic data. Another reason is interconnection of few mitochondrion sections and due to technical limitations. RAZA^{PS} considers such sections as a single object, and calculates its structural parameters from that, and ignores them while searching the objects similar to reference volume.

I have validated the methods developed here using a wide range of samples:

- Synthetic 3D volume: contaminated with different levels of Gaussian and Impulse noise. Simulated tomographic data of GroEL chaperonin.
- High-pressure frozen and freeze-substituted Tomograms: Golgi, mitochondria and insulin secretory granules in a whole cell tomogram reconstructed from a high-pressure frozen, freeze-substituted and plastic-embedded mouse pancreatic beta cell isolated form Islets of Langerhans. Insulin-secreting pancreatic beta cell to compare manual segmentation with semi-automated segmentation. A 3D tomographic reconstruction of a 2D crystal of the V_oV_1 rotary ATPase isolated from *Thermus thermophilus* membranes.
- Cryo-tomographic data: Tomogram from *P.anserina* to detect membrane proteins. *Cryo tomograms of B. bacteriovorus* HD100 cells. Thus I argue the generalization capabilities of our methods on tomographic datasets.

This suggests the tools developed here will be well suited to a wide range of biological datasets. Furthermore, while this thesis primarily aimed to develop automated 3D segmentation algorithms to detect molecules that form part of the photosynthetic machinery of microalgae, the methods are likely to find general applicability for any tomographic data including MRI etc.

Furthermore, images were all recorded using dual axis TEM. The Electron tomography section (1.4.3.3 (Figure 1.22)) summarise the fact that the use of dual axis tomography reduce the lost information from a 'missing wedge' to a smaller 'missing cone'. This is of course a technology constraint not a 3D BLE, RAZA or RAZA^{PS} constraint. However the segmentation algorithm of choice must be robust enough to cope with this. RAZA^{PS}, my algorithm of choice for this purpose, due to its reliance on the second order derivative of voxel intensity, can always generate a complete contour of a segmented 3D object, though noise and the missing cone can affect the precise contour. To test this, cells (Figure 1.3), organelles (Figure 2.5, 3.3, 3.5, 4.7), membrane proteins (Figure 3.4, 3.7, 3.8) and macromolecular assemblies (Figure 2.4, 3.6, 4.3, 4.5) were all segmented and in all cases the segmentation was successful without compensation for the missing wedge. Most notable is the ability of RAZA^{PS} to segment individual ATPases molecules in native membranes (Figure 3.7) to yield contours closely resembling the 3D structure of this complex. This is sufficient for our purposes. Furthermore with the rapid advance of cryo-FIB, the missing wedge/pyramid problem will be reduced further.

The noise models used to test the 3D BLE are provided in section 2.2.2. In summary a range of Gaussian (0-80%) and impulse (0-80%) noise were used (Figure 2.1). In another test (Figure 2.2), the truth sets were contaminated with noise more closely simulating "real" experimental conditions encountered in electron micrographs, in order to provide a more realistic evaluation of the performance of 3D BLE for EM data. . Edge widths as thin as 2 pixels are reproducibly detected with signal intensity and grey scale values as low as 0.72% above the mean of the background noise.

To address this point for the RAZA, GroEL models with a positive control (no noise) and negative control (GroEL reduced in scale by a factor of 2), were noise corrupted using the same Gaussian noise profiles (shown in Figure 4.8).

The experiments were not designed to target a specific SNR. Rather the experiments were designed to identify conditions in which objects could be detected and segmented. The 3D BLE filter was shown to detect 'truth set molecules in noise levels as high as 80% Impulse and 80% Gaussian noise backgrounds. The RAZA filter was shown to detect 'truth set' molecules in 60% Gaussian noise backgrounds (See section 4.3.4)

Through the work in this thesis and that of others we have contributed important proof of principle

data for many of the analytical processes required to construct multi-scale atlases of the *Chlamydomonas reinhardtii* cell. Figure 5.3 below (left to right) shows optical micrographs of *Algae Cells* and an artistically rendered *Algae Cell* section (Courtesy of the Chlamy Centre, www.chlamy.org). *Cell sections* of an algae cell allow the visualisation of organelles, membranes, macromolecules and membrane proteins. To enable automated tomogram segmentation, edge detection software is required to enable high throughput contouring of organelles (e.g. *Mitochondrion*, *Macromolecular assemblies* and *Membrane proteins)*. The *RAZA edge detection* software was developed for this purpose and was shown to be able to detect membrane proteins in artificial membranes (e.g. *3D ATPase segmentation*). Further refinement in sensitivity will be beneficial for the analysis of native membranes imaged in cryo and experiments are underway. Such molecular contours provide constraints into which to dock atomic structures of membrane proteins and macromolecular assemblies. Coupling our *F30* electron microscope with a newly purchased *K2* direct electron detector (with its superior movie mode - *dose fractionation)* allows structure determination towards *~3Å*. This in turn has been interfaced with a powerful supercomputing cluster and ~100 Tb data storage facility. The system was installed late 2015. Standard operating procedures are well underway. Collectively the system is designed to solve atomic resolution structures of isolated membrane proteins using *Single Particle Analysis (SPA)*. These atomic structures can theoretically be docked into the constraining molecular contours in the tomograms with the view to generating 'pseudo-atomic resolution models of the *C. reinhardtii* cells, ultimately to enable structure guided design. The bottom row of panels shows merged data sets for most of the photosynthetic complexes prepared by Janina Steinbeck and Cheng chen Wu, which combine single particle and X-ray crystallography data. A closer look at *Membrane/MP* panel shows circled ATP synthase complex densities into which atomic *ATP synthase structures* could theoretically be fitted. In summary we now have a first multi-scale image-processing pipeline.

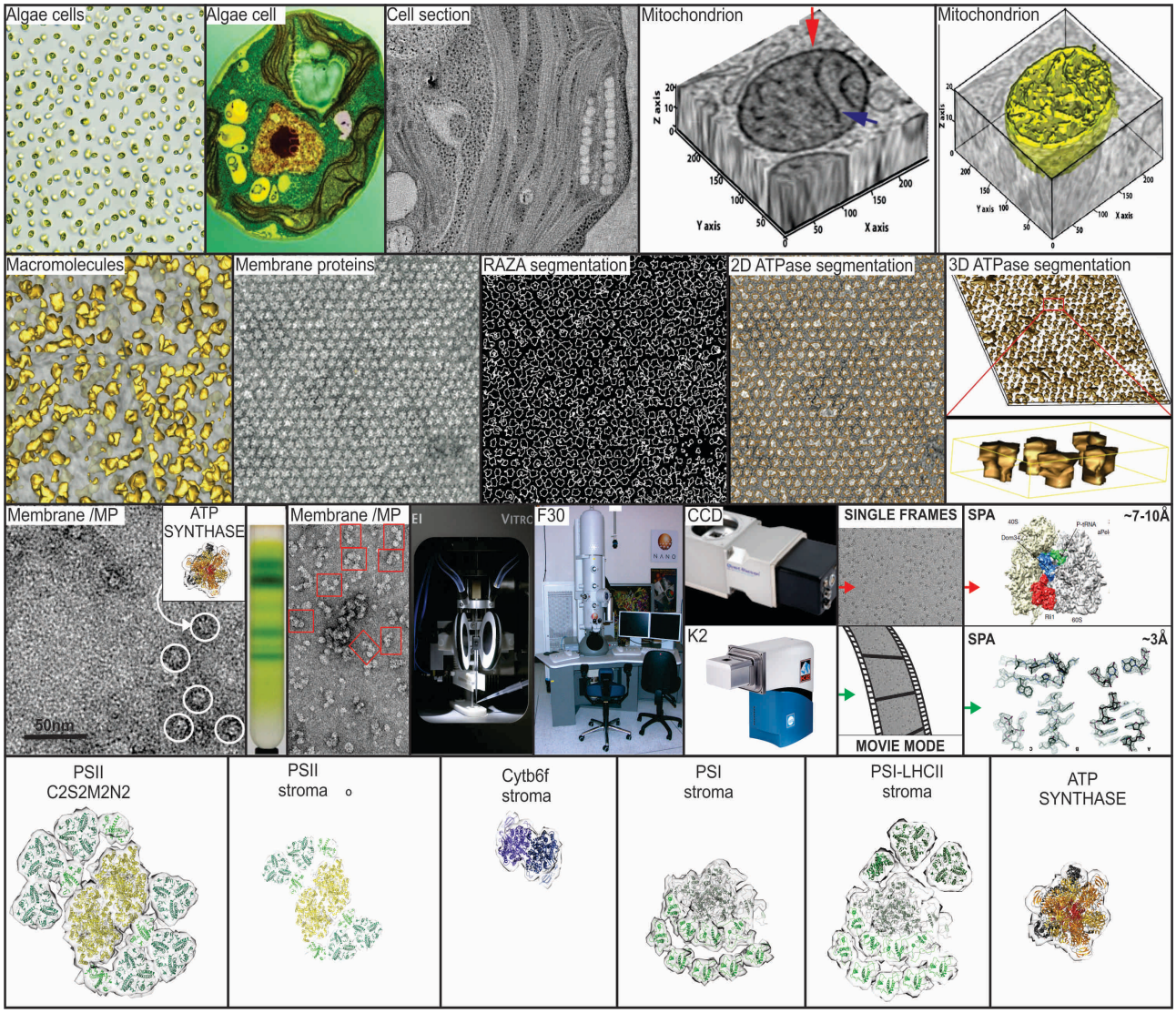


Figure 5.1 Multi-scale image processing pipeline to yield an atomic resolution of photosynthetic machinery of *Chlamydomonas reinhardti*.

5.2 Summary

This thesis described and demonstrated: (1) the limitation of sample preparation and imaging technologies, (2) the advantages and disadvantages of several noise reduction and edge detection techniques, (3) how edge detection assists in segmenting molecules of interest, (4) post-segmentation structure determination, (5) selective molecule segmentation to remove non-related objects, and finally (6) initiating sub-volume averaging and the development of a 3D atlas of macromolecules *in situ*. Figure 5.2 summarises a typical electron microscopy-guided workflow

from sample preparation through to the development of a molecular resolution atlas in 3D.

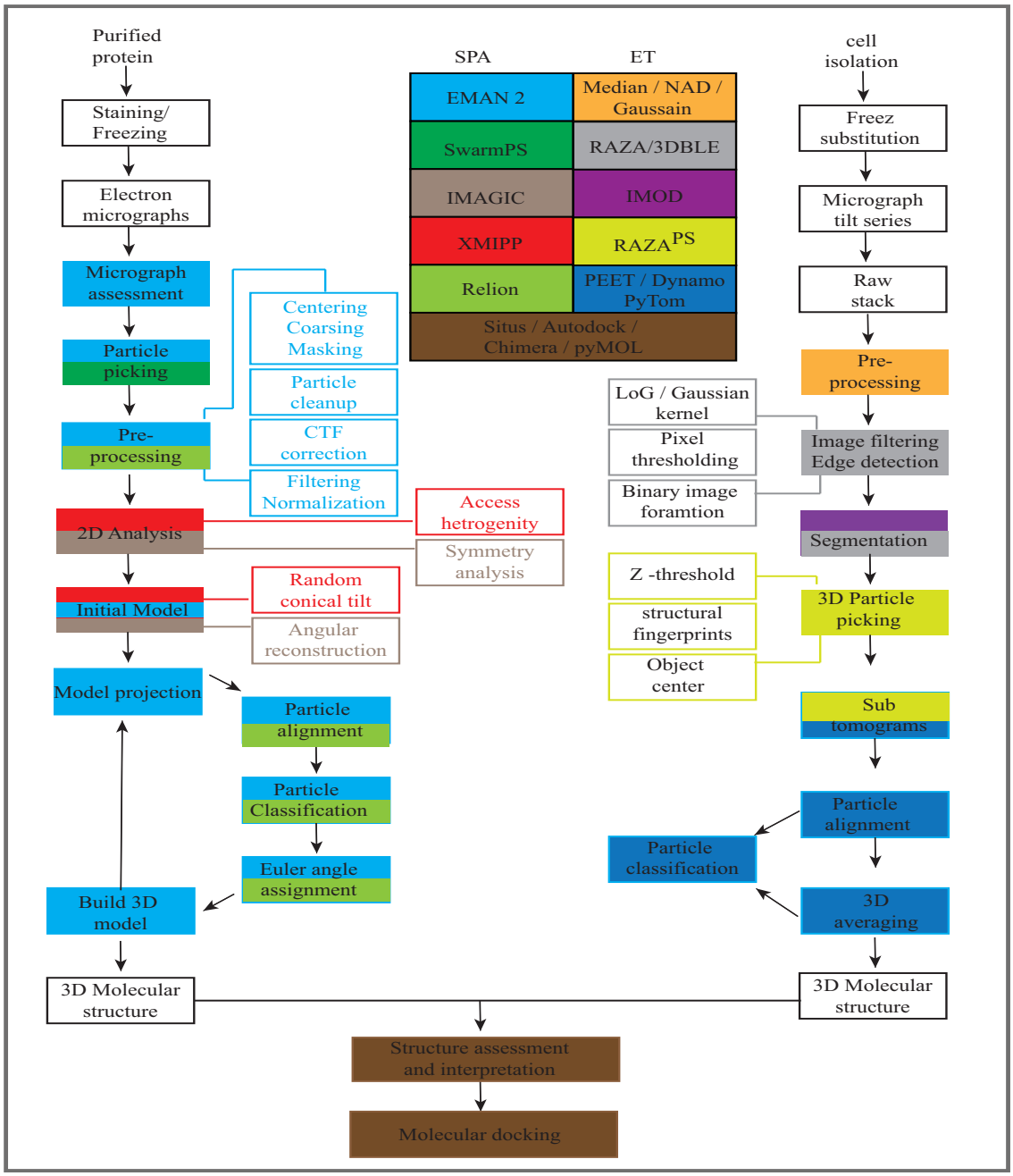


Figure 5.2 Integrating SPA and electron tomographic datasets

The integration of SPA and ET enable molecular docking of 3D structures into electron tomograms. A number of different sequential processes in both techniques have been indicated. The left hand side is an example pipeline of macromolecular structure determination using SPA. The right hand side is a proposed pipeline for 3D structure determination of macromolecules using Electron tomography.

The left hand side of Figure 5.2 shows the methods related to single particle analysis. The right hand side shows the methods that can be used in electron tomography (including the ones developed in this thesis) and how they provide an interface between single particle analysis, single particle tomography and the development of a molecular resolution 3D atlas.

5.3 Future directions

5.3.1 Collection of improved tomograms in single electron counting mode using K2.

Direct electron detection systems and phase plates offer significant improvements to the quality of tomographic datasets. With the reduction of the impacts of sample charging and drift on the captured image, high resolution tomograms can increase the rate of high quality data capture. Single electron counting mode-using Gatan's K2 direct electron camera can, for example, significantly improve the collection of high-quality tomograms and with it the ability to conduct improved edge detection, especially for the analysis of large cellular structures and dynamic macromolecular assemblies that are not easily studied by methods such as X-ray crystallography and NMR spectroscopy.

5.3.2 Development of discriminative bilateral filter to suppress impulse noise further

In this thesis, I developed methods that have helped segment objects of different sizes and shapes. Before the application of the 3D BLE and RAZA methods, I applied Gaussian filtration followed by iterative median filtering to reduce the level of noise. I also used non-linear anisotropic diffusion filtering for suppressing various levels of noise. However, in several examples using these approaches I found that either the edges were not completely detected, or *true* and *spurious* edges were indistinguishable due to high frequency noise in the data. As discussed in Chapter 1, the discriminative bilateral filter (DBL) has a capability to differentiate between *true* edges and high frequency noise because of its inherent photometric exclusion function. However, to the best of my knowledge, only a 2D implementation of DBL is currently available. This could be used to filter tomograms prior to the application of $RAZA^{PS}$, "slice by slice", to see if this yields considerable improvements in pre-processing. Furthermore the DBL filter could be extended to 3D to integrate it with RAZA and suppress high frequency noise more effectively.

5.3.3 Optimize computational resources

Speed: An extensive optimization of methods can reduce the time and space complexity of the algorithms developed further, but this was beyond the scope of this thesis. In Chapter 2 and 3, I show that despite the accuracy and speed of the 3D BLE and RAZA filters compared with manual segmentation, they have higher time and space complexities than other algorithms. Through future work these algorithms could be increased in speed by using efficient algorithms with lower time

complexity, and improved programming. **Graphical User Interface:** The 3D BLE filter was originally developed using the BSoft library, but re-implemented by using IMOD libraries. Now 3D BLE, RAZA and RAZA^{PS} will be available in the IMOD package, but as a command-line tool. A graphical user interface would further improve ease of use for biologists. **Platforms:** Finally, All these developed algorithms have been compiled and tested on Mac OS X operating systems (Snow Leopard). No attempt has been made to test them on Windows and Linux based systems. These methods can be extended to operate on these platforms so that maximum benefits can be achieved.

Based on the capabilities of RAZA, experiments can be initiated on tomograms of individual thylakoid membranes, extended to isolated granal stacks and ultimately to cryo-sections of thylakoid membranes to detect photosystems *in situ*. Should the signal-to-noise ratio of the molecular envelopes be limiting, isolated thylakoid membranes of gold-labelled His-tagged mutants (e.g. PSI, PSII, Cytb_6f and ATPase) can be used to locate the complexes of interest.

5.3.4 High throughput structure determination

After detection and extraction of 3D sub-volumes of these complexes, they could be classified and averaged by using sub-volume averaging programs such as Dynamo (Castano-Diez, Kudryashev et al. 2012). This alignment enhances signal and reduces noise to reveal higher resolution information, improving both structure resolution and docking accuracy. To facilitate high-throughput structure determination and docking of single-particle models into electron tomograms, computational tools for automated 3D tomogram image analysis are also under development.

5.3.5 3D atomic resolution cell atlases by providing segmented contours for docking.

The overall aim of this project was to develop new computational 3D edge detection and particle selection algorithms capable of simultaneous noise suppression and edge detection to initiate the development of a pseudo atomic 3D atlas of the photosynthetic machinery, with the ultimate aim to better enable targeted engineering of the green algae *C. reinhardtii* and optimize the light capture efficiency of microalgal biofuel systems.

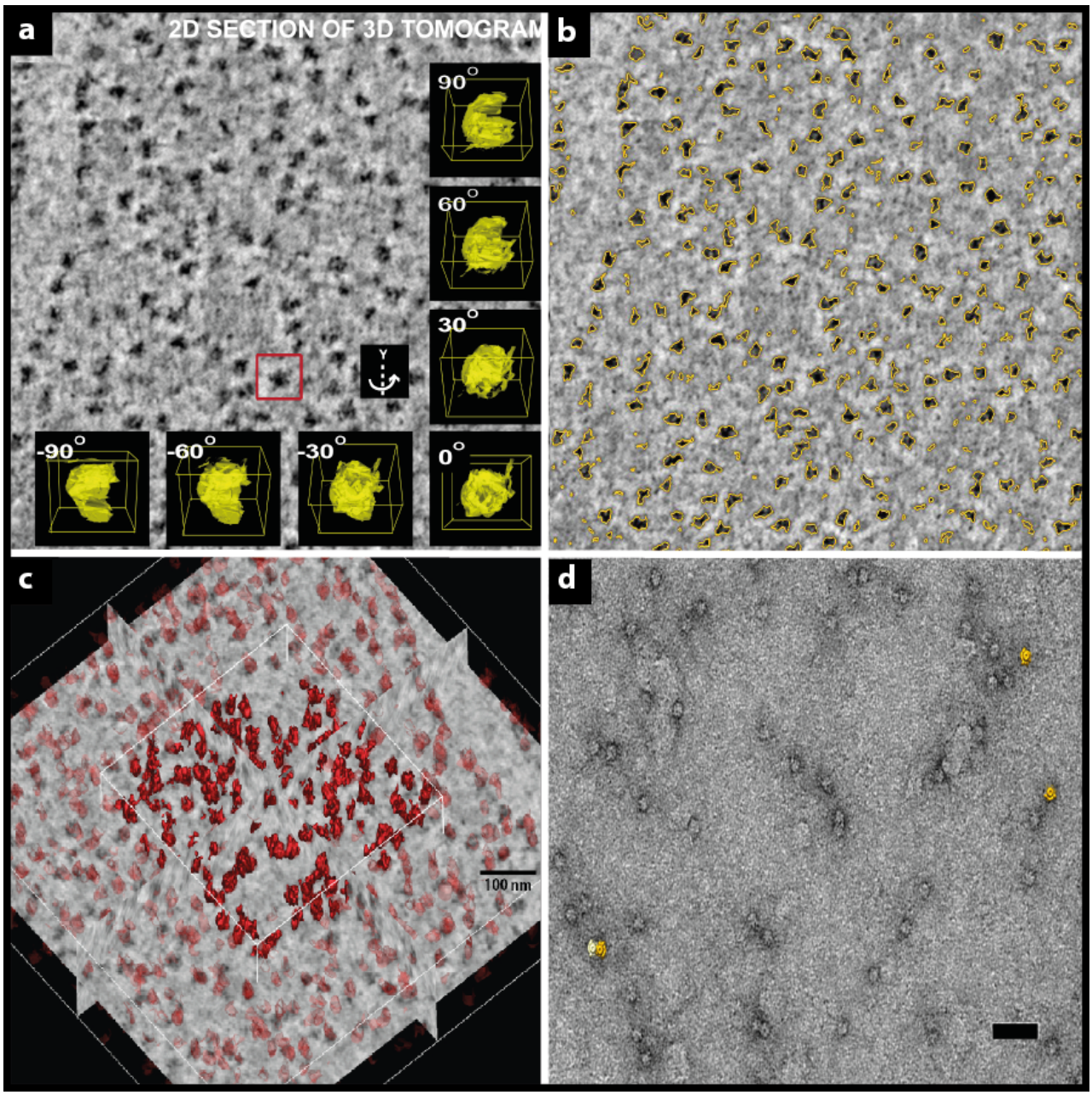


Figure 5.3 Towards automated single particle docking into electron tomogram.

Electron tomogram of chloroplast region of *C.reinhardtii* a). shows the macromolecular assemblies equal to the size of ribosomes. b) Selected density map was then provided to RAZA and RAZA^{PS} for detection of similar objects based on it's '*structural fingerprints*'. RAZA^{PS} then selected all the similar complexes. c) 3D surface view of detected objects. d) ATPases docked in thylakoid membranes of with in a cell of *C. reinhardtii*. Panel 'd' is Image courtesy of J. Steinbeck.

In the preceding chapters I established that my methods could be used to recover molecular-level detail from noisy electron tomograms. In particular I demonstrated that the use of the 3D BLE, RAZA and RAZA^{PS} filters could expedite automated segmentation of macromolecular structures *in situ*. For example, they are able to detect organelles and macromolecules such as ribosomes and extrinsic domains of large membrane protein complexes in membranes. Specifically, in Chapter 3 and 4 I showed that each detected object could be represented by a few mathematical parameters that define its *structural fingerprint*. Furthermore, in Chapter 2 and Chapter 3, by using a

tomographic reconstruction of an insulin-secreting pancreatic beta cell, I showed how results obtained through these methods could be used to initiate the reconstruction of a multi-scale molecular resolution 3D atlas.

Therefore by applying these methods to electron tomograms of the photosynthetic machinery it should be possible to (1) segment proteins of interest by providing reference objects; (2) align similar objects (molecules) and calculate average tomogram structures by using third party software (Castano-Diez, Kudryashev et al. 2012); (3) use these structures to retrospectively dock 3D atomic structures of isolated complexes in public databases such as the Protein Data Bank; and (4) accomplish the nested docking of sub-tomogram averages and fitted atomic coordinates into electron tomograms to develop a molecular resolution atlas.

All key PSII, cyt b_6f, PSI and ATP synthase complexes of the green algae *C. reinhardtii* involved in the state transition process and cyclic electron transport have been resolved (as shown in Figure 1.11). These molecular envelopes allow $RAZA^{PS}$ to provide their segmented contours and calculate their 'molecular fingerprints'. So, a detailed multiscale 3D atlas can be achieved by incorporating details obtained from ET (cellular scale), SPA (molecular scale) and crystallographic data (atomic scale) by docking them back in to the cellular tomogram. As an example, Figure 5.3 (a-c) illustrates high-throughput segmentation with structural fingerprints of ribosome-like particle uncovered by the methods developed in this thesis. These ribosome-like particles are comparable to the size of thylakoid membrane proteins. Figure 5.3d further illustrates the proof of principle through docking of ATPases within the thylakoid membranes. A similar approach can be achieved by docking ribosomes and other molecules involved in photosynthetic process within the *C. reinhardtii* cell. Thus with the help of the computational techniques developed in this thesis, it is feasible to bridge the resolution gap between tomography and atomic resolution techniques capable of solving isolated protein structures on the path to developing pseudo-atomic resolution 3D atlases. Thus we conclude that the achievement of a multiscale structural framework that could guide the engineering of LHC systems is possible with the development of such pipelines for the high throughput determination of large macromolecular structures.

BIBLIOGRAPHY

Abrahams, J. P., A. G. Leslie, R. Lutter and J. E. Walker (1994). "Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria." Nature **370**(6491): 621-628.

Adiga, P. S., R. Malladi, W. Baxter and R. M. Glaeser (2004). "A binary segmentation approach for boxing ribosome particles in cryo EM micrographs." J Struct Biol **145**(1-2): 142-151.

Adiga, U., W. T. Baxter, R. J. Hall, B. Rockel, B. K. Rath, J. Frank and R. Glaeser (2005). "Particle picking by segmentation: a comparative study with SPIDER-based manual particle picking." J Struct Biol **152**(3): 211-220.

Alber, F., F. Forster, D. Korkin, M. Topf and A. Sali (2008). "Integrating diverse data for structure determination of macromolecular assemblies." Annu Rev Biochem **77**: 443-477.

Ali, R. A., M. J. Landsberg, E. Knauth, G. P. Morgan, B. J. Marsh and B. Hankamer (2012). "A 3D image filter for parameter-free segmentation of macromolecular structures from electron tomograms." PLoS One **7**(3): e33697.

Allegretti, M., D. J. Mills, G. McMullan, W. Kuhlbrandt and J. Vonck (2014). "Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector." Elife **3**: e01963.

Amat, F., F. Moussavi, L. R. Comolli, G. Elidan, K. H. Downing and M. Horowitz (2008). "Markov random field based automatic image alignment for electron tomography." J Struct Biol **161**(3): 260-275.

Amunts, A., A. Brown, X. C. Bai, J. L. Llacer, T. Hussain, P. Emsley, F. Long, G. Murshudov, S. H. Scheres and V. Ramakrishnan (2014). "Structure of the yeast mitochondrial large ribosomal subunit." Science **343**(6178): 1485-1489.

Amunts, A., O. Drory and N. Nelson (2007). "The structure of a plant photosystem I supercomplex at 3.4 A resolution." Nature **447**(7140): 58-63.

Amunts, A., H. Toporik, A. Borovikova and N. Nelson (2010). "Structure determination and improved model of plant photosystem I." J Biol Chem **285**(5): 3478-3486.

Asano, S., Y. Fukuda, F. Beck, A. Aufderheide, F. Forster, R. Danev and W. Baumeister (2015). "Proteasomes. A molecular census of 26S proteasomes in intact neurons." Science **347**(6220): 439-442.

Bajaj, C., Z. Yu and M. Auer (2003). "Volumetric feature extraction and visualization of tomographic molecular imaging." J Struct Biol **144**(1-2): 132-143.

Baker, M. L., Z. Yu, W. Chiu and C. Bajaj (2006). "Automated segmentation of molecular subunits in electron cryomicroscopy density maps." J Struct Biol **156**(3): 432-441.

Ban, N., P. Nissen, J. Hansen, P. B. Moore and T. A. Steitz (2000). "The complete atomic structure of the large ribosomal subunit at 2.4 A resolution." Science **289**(5481): 905-920.

Baniulis, D., E. Yamashita, J. P. Whitelegge, A. I. Zatsman, M. P. Hendrich, S. S. Hasan, C. M. Ryan and W. A. Cramer (2009). "Structure-Function, Stability, and Chemical Modification of the Cyanobacterial Cytochrome b6f Complex from Nostoc sp. PCC 7120." J Biol Chem **284**(15): 9861-9869.

Barber, J., J. Nield, E. Morris, D. Zheleva and B. Hankamer (1997). "The structure, function and dynamics of photosystem two." Physiologia plantarum **100**(4): 817-827.

Bartesaghi, A., D. Matthies, S. Banerjee, A. Merk and S. Subramaniam (2014). "Structure of beta-galactosidase at 3.2-A resolution obtained by cryo-electron microscopy." Proc Natl Acad Sci U S A **111**(32): 11709-11714.

Bartesaghi, A. and S. Subramaniam (2009). "Membrane protein structure determination using cryo-electron tomography and 3D image averaging." Curr Opin Struct Biol **19**(4): 402-407.

Baumeister, W. (2002). "Electron tomography: towards visualizing the molecular organization of the cytoplasm." Curr Opin Struct Biol **12**(5): 679-684.

Baumeister, W., R. Grimm and J. Walz (1999). "Electron tomography of molecules and cells." Trends Cell Biol **9**(2): 81-85.

Berggren, G., A. Adamska, C. Lambertz, T. R. Simmons, J. Esselborn, M. Atta, S. Gambarelli, J. M. Mouesca, E. Reijerse, W. Lubitz, T. Happe, V. Artero and M. Fontecave (2013). "Biomimetic assembly and activation of [FeFe]-hydrogenases." Nature **499**(7456): 66-69.

Block, M. A., A.-J. Dorne, J. Joyard and R. Douce (1983). "Preparation and characterization of membrane fractions enriched in outer and inner envelope membranes from spinach chloroplasts. II. Biochemical characterization." Journal of Biological Chemistry **258**(21): 13281-13286.

Boekema, E. J., B. Hankamer, D. Bald, J. Kruip, J. Nield, A. F. Boonstra, J. Barber and M. Rogner (1995). "Supramolecular structure of the photosystem II complex from green plants and cyanobacteria." Proc Natl Acad Sci U S A **92**(1): 175-179.

Bohm, J., A. S. Frangakis, R. Hegerl, S. Nickell, D. Typke and W. Baumeister (2000). "Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms." Proc Natl Acad Sci U S A **97**(26): 14245-14250.

Boyer, P. and J. Walker. (1997). "Press Release: The 1997 Nobel Prize in Chemistry." Retrieved 03/06/2015, from <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1997/press.html>.

Brandt, F., S. A. Etchells, J. O. Ortiz, A. H. Elcock, F. U. Hartl and W. Baumeister (2009). "The native 3D organization of bacterial polysomes." Cell **136**(2): 261-271.

Brimblecombe, R., G. C. Dismukes, G. F. Swiegers and L. Spiccia (2009). "Molecular water-oxidation catalysts for photoelectrochemical cells." Dalton Trans(43): 9374-9384.

Bureau, U. S. C. " Population Division, International database." from http://www.census.gov/ipc/www/idb/worldpopgraph.php.

Campbell, M. G., D. Veesler, A. Cheng, C. S. Potter and B. Carragher (2015). "2.8 A resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy." Elife **4**.

Canny, J. (1986). "A computational Approach to Edge Detection." IEEE Trans. Pattern Analysis and Machine Intelligence **8**(6): 19.

Canny, J. (1986). "A computational approach to Edge detection." IEEE Transactions on Pattern analysis and Machine Intelligence **PAMI-8**(6): 679-698.

Castano-Diez, D., M. Kudryashev, M. Arheit and H. Stahlberg (2012). "Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments." J Struct Biol **178**(2): 139-151.

Catarious, D. M., Jr., A. H. Baydush and C. E. Floyd, Jr. (2006). "Characterization of difference of Gaussian filters in the detection of mammographic regions." Med Phys **33**(11): 4104-4114.

Chandramouli, P., M. Topf, J. F. Menetret, N. Eswar, J. J. Cannone, R. R. Gutell, A. Sali and C. W. Akey (2008). "Structure of the mammalian 80S ribosome at 8.7 A resolution." Structure **16**(4): 535-548.

Chari, A., D. Haselbach, J. M. Kirves, J. Ohmer, E. Paknia, N. Fischer, O. Ganichkin, V. Moller, J. J. Frye, G. Petzold, M. Jarvis, M. Tietzel, C. Grimm, J. M. Peters, B. A. Schulman, K. Tittmann, J. Markl, U. Fischer and H. Stark (2015). "ProteoPlex: stability optimization of macromolecular complexes by sparse-matrix screening of chemical space." Nat Methods **12**(9): 859-865.

Cong, Y. and S. J. Ludtke (2010). "Single particle analysis at high resolution." Methods Enzymol **482**: 211-235.

Cope, J., S. Gilbert, I. Rayment, D. Mastronarde and A. Hoenger (2010). "Cryo-electron tomography of microtubule-kinesin motor complexes." J Struct Biol **170**(2): 257-265.

Cyrklaff, M., C. Risco, J. J. Fernandez, M. V. Jimenez, M. Esteban, W. Baumeister and J. L. Carrascosa (2005). "Cryo-electron tomography of vaccinia virus." Proc Natl Acad Sci U S A **102**(8): 2772-2777.

Danev, R. and W. Baumeister (2016). "Cryo-EM single particle analysis with the Volta phase plate." Elife **5**.

Danev, R., B. Buijsse, M. Khoshouei, J. M. Plitzko and W. Baumeister (2014). "Volta potential phase plate for in-focus phase contrast transmission electron microscopy." Proc Natl Acad Sci U S A **111**(44): 15635-15640.

Daum, B., D. Nicastro, J. Austin, 2nd, J. R. McIntosh and W. Kuhlbrandt (2010). "Arrangement of photosystem II and ATP synthase in chloroplast membranes of spinach and pea." Plant Cell **22**(4): 1299-1312.

Davies, K. M., C. Anselmi, I. Wittig, J. D. Faraldo-Gomez and W. Kuhlbrandt (2012). "Structure of the yeast F1Fo-ATP synthase dimer and its role in shaping the mitochondrial cristae." Proc Natl Acad Sci U S A **109**(34): 13602-13607.

Davies, K. M., M. Strauss, B. Daum, J. H. Kief, H. D. Osiewacz, A. Rycovska, V. Zickermann and W. Kuhlbrandt (2011). "Macromolecular organization of ATP synthase and complex I in whole mitochondria." Proc Natl Acad Sci U S A **108**(34): 14121-14126.

Dekker, J. P. and E. J. Boekema (2005). "Supramolecular organization of thylakoid membrane proteins in green plants." Biochim Biophys Acta **1706**(1-2): 12-39.

Denk, W. and H. Horstmann (2004). "Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure." PLoS Biol **2**(11): e329.

Deriche, D., J. Cocquerez and C. Almouzyn (1988). An efficient method to build early image description. Pattern recognition. Rome, Italy.

Dinca, M., Y. Surendranath and D. G. Nocera (2010). "Nickel-borate oxygen-evolving catalyst that functions under benign conditions." Proc Natl Acad Sci U S A **107**(23): 10337-10341.

Dittami, S. M., G. Michel, J. Collen, C. Boyen and T. Tonon (2010). "Chlorophyll-binding proteins revisited--a multigenic family of light-harvesting and stress proteins from a brown algal perspective." BMC Evol Biol **10**: 365.

Drop, B., M. Webber-Birungi, F. Fusetti, R. Kouril, K. E. Redding, E. J. Boekema and R. Croce (2011). "Photosystem I of Chlamydomonas reinhardtii contains nine light-harvesting complexes (Lhca) located on one side of the core." J Biol Chem **286**(52): 44878-44887.

Drop, B., M. Webber-Birungi, S. K. Yadav, A. Filipowicz-Szymanska, F. Fusetti, E. J. Boekema and R. Croce (2014). "Light-harvesting complex II (LHCII) and its supramolecular organization in Chlamydomonas reinhardtii." Biochim Biophys Acta **1837**(1): 63-72.

Engel, B. D., M. Schaffer, L. Kuhn Cuellar, E. Villa, J. M. Plitzko and W. Baumeister (2015). "Native architecture of the Chlamydomonas chloroplast revealed by in situ cryo-electron tomography." Elife **4**.

FAO, U. (2009). How to Feed the World in 2050. Rome: High-Level Expert Forum.

Fernandez, J. J. (2009). "TOMOBFLOW: feature-preserving noise filtering for electron tomography." BMC Bioinformatics **10**: 178.

Fernandez, J. J. and J. L. Carrascosa (2010). "Image processing in electron tomography." Microscopy **Science, Technology, Application and Education**.

Fernandez, J. J., D. Luque, J. R. Caston and J. L. Carrascosa (2008). "Sharpening high resolution information in single particle electron cryomicroscopy." J Struct Biol **164**(1): 170-175.

Ferreira, K. N., T. M. Iverson, K. Maghlaoui, J. Barber and S. Iwata (2004). "Architecture of the photosynthetic oxygen-evolving center." Science **303**(5665): 1831-1838.

Fischer, N., P. Neumann, A. L. Konevega, L. V. Bock, R. Ficner, M. V. Rodnina and H. Stark (2015). "Structure of the E. coli ribosome-EF-Tu complex at <3 A resolution by Cs-corrected cryo-EM." Nature **520**(7548): 567-570.

Forster, F., B. G. Han and M. Beck (2010). "Visual proteomics." Methods Enzymol **483**: 215-243.

Forster, F. and R. Hegerl (2007). "Structure determination in situ by averaging of tomograms." Methods Cell Biol **79**: 741-767.

Forster, F., S. Pruggnaller, A. Seybert and A. S. Frangakis (2008). "Classification of cryo-electron sub-tomograms using constrained correlation." J Struct Biol **161**(3): 276-286.

Frangakis, A. S., J. Bohm, F. Forster, S. Nickell, D. Nicastro, D. Typke, R. Hegerl and W. Baumeister (2002). "Identification of macromolecular complexes in cryoelectron tomograms of phantom cells." Proc Natl Acad Sci U S A **99**(22): 14153-14158.

Frangakis, A. S. and F. Forster (2004). "Computational exploration of structural information from cryo-electron tomograms." Curr Opin Struct Biol **14**(3): 325-331.

Frangakis, A. S. and R. Hegerl (2001). "Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion." J Struct Biol **135**(3): 239-250.

Frangakis, A. S. and R. Hegerl (2002). "Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis." J Struct Biol **138**(1-2): 105-113.

Gaffron, H. (1939). "Reduction of Carbon Dioxide with Molecular Hydrogen in Gree Algae." Nature **143**: 204-205.

Gan, L. and G. J. Jensen (2012). "Electron tomography of cells." Quarterly reviews of biophysics **45**(01): 27-56.

Garduno, E., M. Wong-Barnum, N. Volkmann and M. H. Ellisman (2008). "Segmentation of electron tomographic data sets using fuzzy set theory principles." J Struct Biol **162**(3): 368-379.

Gerle, C., K. Tani, K. Yokoyama, M. Tamakoshi, M. Yoshida, Y. Fujiyoshi and K. Mitsuoka (2006). "Two-dimensional crystallization and analysis of projection images of intact Thermus thermophilus V-ATPase." J Struct Biol **153**(2): 200-206.

Gibbons, C., M. G. Montgomery, A. G. Leslie and J. E. Walker (2000). "The structure of the central stalk in bovine F(1)-ATPase at 2.4 A resolution." Nat Struct Biol **7**(11): 1055-1061.

Giordano, M., J. Beardall and J. A. Raven (2005). "CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution." Annu Rev Plant Biol **56**: 99-131.

Gold, V. A., R. Ieva, A. Walter, N. Pfanner, M. van der Laan and W. Kuhlbrandt (2014). "Visualizing active membrane protein complexes by electron cryotomography." Nat Commun **5**: 4129.

Gonzalez, R. C. (2002). Digital Image Processing, Prentice Hall, New Jersey.

Greenwell, H. C., L. M. Laurens, R. J. Shields, R. W. Lovitt and K. J. Flynn (2010). "Placing microalgae on the biofuels priority list: a review of the technological challenges." J R Soc Interface **7**(46): 703-726.

Grigorieff, N. (2007). "FREALIGN: high-resolution refinement of single particle structures." J Struct Biol **157**(1): 117-125.

Groth, G. and E. Pohl (2001). "The structure of the chloroplast F1-ATPase at 3.2 A resolution." J Biol Chem **276**(2): 1345-1352.

Gupta, S. K., S. Kumari, K. Reddy and F. Bux (2013). "Trends in biohydrogen production: major challenges and state-of-the-art developments." Environ Technol **34**(13-16): 1653-1670.

Guskov, A., J. Kern, A. Gabdulkhakov, M. Broser, A. Zouni and W. Saenger (2009). "Cyanobacterial photosystem II at 2.9-A resolution and the role of quinones, lipids, channels and chloride." Nat Struct Mol Biol **16**(3): 334-342.

Hankamer, B., F. Lehr, J. Rupprecht, J. H. Mussgnug, C. Posten and O. Kruse (2007). "Photosynthetic biomass and H2 production by green algae: from bioengineering to bioreactor scale-up." Physiol Plant **131**(1): 10-21.

Hankamer, B., E. Morris, J. Nield, C. Gerle and J. Barber (2001). "Three-dimensional structure of the photosystem II core dimer of higher plants determined by electron microscopy." J Struct Biol **135**(3): 262-269.

Hankamer, B., E. P. Morris and J. Barber (1999). "Revealing the structure of the oxygen-evolving core dimer of photosystem II by cryoelectron crystallography." Nat Struct Biol **6**(6): 560-564.

Harris, E. H. (2009). The Chlamydomonas sourcebook: introduction to Chlamydomonas and its laboratory use, Academic Press.

Hegemann, P. (1997). "Vision in microalgae." Planta **203**(3): 265-274.

Heldt, H. W. and F. Sauer (1971). "The inner membrane of the chloroplast envelope as the site of specific metabolite transport." Biochim Biophys Acta **234**(1): 83-91.

Hellmich, U. A. and R. Gaudet (2014). "Structural biology of TRP channels." Handb Exp Pharmacol **223**: 963-990.

Henderson, R. (2013). "Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise." Proc Natl Acad Sci U S A **110**(45): 18037-18041.

Heymann, J. B. and D. M. Belnap (2007). "Bsoft: image processing and molecular modeling for electron microscopy." J Struct Biol **157**(1): 3-18.

Heymann, J. B., G. Cardone, D. C. Winkler and A. C. Steven (2008). "Computational resources for cryo-electron tomography in Bsoft." J Struct Biol **161**(3): 232-242.

Hohn, M., G. Tang, G. Goodyear, P. R. Baldwin, Z. Huang, P. A. Penczek, C. Yang, R. M. Glaeser, P. D. Adams and S. J. Ludtke (2007). "SPARX, a new environment for Cryo-EM image processing." J Struct Biol **157**(1): 47-55.

Hrabe, T., Y. Chen, S. Pfeffer, L. K. Cuellar, A. V. Mangold and F. Forster (2012). "PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis." J Struct Biol **178**(2): 177-188.

Hrabe, T. and F. Förster (2011). Structure Determination by Single Particle Cryo-electron Tomography, John Wiley & Sons, Ltd.

Hunter, P. (2010). "The tide turns towards microalgae. Current research aims to produce traditional biofuels from algae, but their potential to generate sustainable energy might be even greater and more 'natural'." EMBO Rep **11**(8): 583-586.

IEA (2010). World energy outlook. Paris, IEA.

Inoue, K. (2007). "The chloroplast outer envelope membrane: the edge of light and excitement." Journal of Integrative Plant Biology **49**(8): 1100-1111.

IPCC. (2014). "Climate Change." from http://www.ipcc.ch/report/ar5/ (accessed November 2014).

Iwai, M., M. Yokono, N. Inada and J. Minagawa (2010). "Live-cell imaging of photosystem II antenna dissociation during state transitions." Proc Natl Acad Sci U S A **107**(5): 2337-2342.

Iwasaki, K. (2006). "Single-particle analysis and electron tomography." International Journal of Nanotechnology **3**(4): 480-491.

Jiang, W., M. L. Baker, Q. Wu, C. Bajaj and W. Chiu (2003). "Applications of a bilateral denoising filter in biological electron microscopy." J Struct Biol **144**(1-2): 114-122.

Jimenez, N., K. Vocking, E. G. van Donselaar, B. M. Humbel, J. A. Post and A. J. Verkleij (2009). "Tannic acid-mediated osmium impregnation after freeze-substitution: a strategy to enhance membrane contrast for electron tomography." J Struct Biol **166**(1): 103-106.

Jin, Y. (2006). Multi-objective machine learning.

John, C. (1986). "A Computational Approach to Edge Detection." IEEE Transactions: Pattern Analysis and Machine Intelligence **PAMI-8**(6): 679-698.

Judith M. S. Prewitt (1970). "Object Enhancement and Extraction." Picture processing and Psychopictorics",Academic Press.

Kargul, J., M. V. Turkina, J. Nield, S. Benson, A. V. Vener and J. Barber (2005). "Light-harvesting complex II protein CP29 binds to photosystem I of Chlamydomonas reinhardtii under State 2 conditions." FEBS J **272**(18): 4797-4806.

Kass, M., A. Witkin and D. Terzopoulos (1988). "Snakes: Active contour models." International journal of computer vision: 321-331.

Kastner, B., N. Fischer, M. M. Golas, B. Sander, P. Dube, D. Boehringer, K. Hartmuth, J. Deckert, F. Hauer, E. Wolf, H. Uchtenhagen, H. Urlaub, F. Herzog, J. M. Peters, D. Poerschke, R. Luhrmann and H. Stark (2008). "GraFix: sample preparation for single-particle electron cryomicroscopy." Nat Methods **5**(1): 53-55.

Kong, H., H. C. Akakin and S. E. Sarma (2013). "A generalized Laplacian of Gaussian filter for blob detection and its applications." IEEE Trans Cybern **43**(6): 1719-1733.

Koster, A. J., R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz and W. Baumeister (1997). "Perspectives of molecular and cellular electron tomography." J Struct Biol **120**(3): 276-308.

Kouril, R., J. P. Dekker and E. J. Boekema (2012). "Supramolecular organization of photosystem II in green plants." Biochim Biophys Acta **1817**(1): 2-12.

Kreimer, G. (2009). "The green algal eyespot apparatus: a primordial visual system and more?" Curr Genet **55**(1): 19-43.

Kremer, J. R., D. N. Mastronarde and J. R. McIntosh (1996). "Computer visualization of three-dimensional image data using IMOD." J Struct Biol **116**(1): 71-76.

Kuhlbrandt, W. (2014). "Biochemistry. The resolution revolution." Science **343**(6178): 1443-1444.

Kuhlbrandt, W. (2014). "Cryo-EM enters a new era." Elife **3**: e03678.

Lambert, C. and R. E. Sockett (2005). Laboratory Maintenance of Bdellovibrio. Current Protocols in Microbiology, John Wiley & Sons, Inc.

Landsberg, M. J., S. A. Jones, R. Rothnagel, J. N. Busby, S. D. Marshall, R. M. Simpson, J. S. Lott, B. Hankamer and M. R. Hurst (2011). "3D structure of the Yersinia entomophaga toxin complex and implications for insecticidal activity." Proceedings of the National Academy of Sciences **108**(51): 20544-20549.

Lawrence, A., J. C. Bouwer, G. Perkins and M. H. Ellisman (2006). "Transform-based backprojection for volume reconstruction of large format electron microscope tilt series." J Struct Biol **154**(2): 144-167.

Lebbink, M. N., W. J. Geerts, T. P. van der Krift, M. Bouwhuis, L. O. Hertzberger, A. J. Verkleij and A. J. Koster (2007). "Template matching as a tool for annotation of tomograms of stained biological structures." J Struct Biol **158**(3): 327-335.

Lebbink, M. N., E. van Donselaar, B. M. Humbel, L. O. Hertzberger, J. A. Post and A. J. Verkleij (2009). "Induced membrane domains as visualized by electron tomography and template matching." J Struct Biol **166**(2): 156-161.

Lebeau, L. and C. Venien-Bryan (2013). "Monolayer two-dimensional crystallization of membrane proteins." Methods Mol Biol **955**: 59-71.

Li, X., P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard and Y. Cheng (2013). "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM." Nat Methods **10**(6): 584-590.

Liao, M., E. Cao, D. Julius and Y. Cheng (2013). "Structure of the TRPV1 ion channel determined by electron cryo-microscopy." Nature **504**(7478): 107-112.

Liu, J., E. R. Wright and H. Winkler (2010). "3D visualization of HIV virions by cryoelectron tomography." Methods Enzymol **483**: 267-290.

Lu, P., X. C. Bai, D. Ma, T. Xie, C. Yan, L. Sun, G. Yang, Y. Zhao, R. Zhou, S. H. Scheres and Y. Shi (2014). "Three-dimensional structure of human gamma-secretase." Nature **512**(7513): 166-170.

Lucic, V., F. Forster and W. Baumeister (2005). "Structural studies by electron tomography: from cells to molecules." Annu Rev Biochem **74**: 833-865.

Ludtke, S. J., D. H. Chen, J. L. Song, D. T. Chuang and W. Chiu (2004). "Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy." Structure **12**(7): 1129-1136.

Lyumkis, D., J. P. Julien, N. de Val, A. Cupo, C. S. Potter, P. J. Klasse, D. R. Burton, R. W. Sanders, J. P. Moore, B. Carragher, I. A. Wilson and A. B. Ward (2013). "Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer." Science **342**(6165): 1484-1490.

Mallick, S. P., Y. Zhu and D. Kriegman (2004). "Detecting particles in cryo-EM micrographs using learned features." J Struct Biol **145**(1-2): 52-62.

Marr, D. and E. Hildreth (1980). "Theory of edge detection." Proceedings of the Royal Society of London. Series B. Biological Sciences **207**(1167): 187-217.

Marr, D. and E. Hildreth (1980). "Theory of edge detection." Proc R Soc Lond B Biol Sci **207**(1167): 187-217.

Marsh, B. J., D. N. Mastronarde, K. F. Buttle, K. E. Howell and J. R. McIntosh (2001). "Organellar relationships in the Golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography." Proc Natl Acad Sci U S A **98**(5): 2399-2406.

Marsh, B. J., D. N. Mastronarde, K. F. Buttle, K. E. Howell and J. R. McIntosh (2001). "Organellar relationships in the Golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography." Proceedings of the National Academy of Sciences of the United States of America **98**(5): 2399-2406.

Marsh, B. J., N. Volkmann, J. R. McIntosh and K. E. Howell (2004). "Direct continuities between cisternae at different levels of the Golgi complex in glucose-stimulated mouse islet beta cells." Proc Natl Acad Sci U S A **101**(15): 5565-5570.

Martinez-Sanchez, A., I. Garcia, S. Asano, V. Lucic and J. J. Fernandez (2014). "Robust membrane detection based on tensor voting for electron tomography." J Struct Biol **186**(1): 49-61.

Martinez-Sanchez, A., I. Garcia and J.-J. Fernandez (2013). "A ridge-based framework for segmentation of 3D electron microscopy datasets." Journal of Structural Biology **181**(1): 61-70.

Martinez-Sanchez, A., I. Garcia and J. J. Fernandez (2011). "A differential structure approach to membrane segmentation in electron tomography." J Struct Biol **175**(3): 372-383.

Mastronarde, D. N. (1997). "Dual-axis tomography: an approach with alignment methods that preserve resolution." J Struct Biol **120**(3): 343-352.

Mastronarde, D. N. (2005). "Automated electron microscope tomography using robust prediction of specimen movements." J Struct Biol **152**(1): 36-51.

Mazor, Y., A. Borovikova, I. Greenberg and N. Nelson (2015). Crystal structure of plant Photosystem I at 3.1 Angstrom resolution.

McMullan, G., S. Chen, R. Henderson and A. R. Faruqi (2009). "Detective quantum efficiency of electron area detectors in electron microscopy." Ultramicroscopy **109**(9): 1126-1143.

McMullan, G., A. R. Faruqi, D. Clare and R. Henderson (2014). "Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy." Ultramicroscopy **147**: 156-163.

Mescher, A. (2009). Junqueira's Basic Histology: Text and Atlas, 12th Edition : Text and Atlas: Text and Atlas, Mcgraw-hill.

Messaoudii, C., T. Boudier, C. O. Sanchez Sorzano and S. Marco (2007). "TomoJ: tomography software for three-dimensional reconstruction in transmission electron microscopy." BMC Bioinformatics **8**: 288.

Mitra, M. and A. Melis (2008). "Optical properties of microalgae for enhanced biofuels production." Opt Express **16**(26): 21807-21820.

Monga, O., R. Deriche, G. Malandain and J. P. Cocquerez (1991). "Recursive filtering and edge tracking: two primary tools for 3D edge detection." Image and vision computing **9**(4): 203-214.

Monga, O. D., R. Malandain,G. Cocquerez, J, P. (1991). "Recursive filtering and edge tracking: two primary tools and 3D edge detection." Image and Vision Computing **9**(4): 203-214.

Morita, E., H. Kuroiwa, T. Kuroiwa and H. Nozaki (1997). "HIGH LOCALIZATION OF RIBULOSE□1, 5□BISPHOSPHATE CARBOXULASE/OXYGENASE IN THE PYRENOIDS OF CHLAMYDOMONAS REINHARDTII (CHLOROPHYTA), AS REVEALED BY CRYOFIXATION AND IMMUNOGOLD ELECTRON MICROSCOPY." Journal of phycology **33**(1): 68-72.

Moriya, T., E. Acar, R. H. Cheng and U. Ruotsalainen (2015). "A Bayesian approach for suppression of limited angular sampling artifacts in single particle 3D reconstruction." J Struct Biol **191**(3): 318-331.

Mukherjee, B. and J. V. Moroney "Algal Carbon Dioxide Concentrating Mechanisms." eLS.

Munekage, Y., M. Hashimoto, C. Miyake, K. Tomizawa, T. Endo, M. Tasaka and T. Shikanai (2004). "Cyclic electron flow around photosystem I is essential for photosynthesis." Nature **429**(6991): 579-582.

Murata, N. (1969). "Control of excitation transfer in photosynthesis. I. Light-induced change of chlorophyll a fluorescence in Porphyridium cruentum." Biochim Biophys Acta **172**(2): 242-251.

Mussgnug, J. H., S. Thomas-Hall, J. Rupprecht, A. Foo, V. Klassen, A. McDowall, P. M. Schenk, O. Kruse and B. Hankamer (2007). "Engineering photosynthetic light capture: impacts on improved solar energy to biomass conversion." Plant Biotechnol J **5**(6): 802-814.

Mustardy, L. and G. Garab (2003). "Granum revisited. A three-dimensional model--where things fall into place." Trends Plant Sci **8**(3): 117-122.

Nagayoshi, M., K. Murase, K. Fujino, Y. Uenishi, M. Kawamata, Y. Nakamura, K. Kitamura, I. Higuchi, N. Oku and J. Hatazawa (2005). "Usefulness of noise adaptive non-linear gaussian filter in FDG-PET study." Ann Nucl Med **19**(6): 469-477.

Nicastro, D., C. Schwartz, J. Pierson, R. Gaudette, M. E. Porter and J. R. McIntosh (2006). "The molecular architecture of axonemes revealed by cryoelectron tomography." Science **313**(5789): 944-948.

Nickell, S., F. Forster, A. Linaroudis, W. D. Net, F. Beck, R. Hegerl, W. Baumeister and J. M. Plitzko (2005). "TOM software toolbox: acquisition and analysis for electron tomography." J Struct Biol **149**(3): 227-234.

Nickell, S., C. Kofler, A. P. Leis and W. Baumeister (2006). "A visual approach to proteomics." Nat Rev Mol Cell Biol **7**(3): 225-230.

Noske, A. B., A. J. Costin, G. P. Morgan and B. J. Marsh (2008). "Expedited approaches to whole cell electron tomography and organelle mark-up in situ in high-pressure frozen pancreatic islets." J Struct Biol **161**(3): 298-313.

OECD (2014). OfEC-oaD Environmental outlook to 2050, Organisation for Economic Co-operation and Development.

Ortiz, J. O., F. Forster, J. Kurner, A. A. Linaroudis and W. Baumeister (2006). "Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition." J Struct Biol **156**(2): 334-341.

Oster, G. and H. Wang (2000). "Reverse engineering a protein: the mechanochemistry of ATP synthase." Biochimica et Biophysica Acta (BBA)-Bioenergetics **1458**(2): 482-510.

Paavolainen, L., E. Acar, U. Tuna, S. Peltonen, T. Moriya, P. Soonsawad, V. Marjomaki, R. H. Cheng and U. Ruotsalainen (2014). "Compensation of missing wedge effects with sequential statistical reconstruction in electron tomography." PLoS One **9**(10): e108978.

Pantelic, R. S., G. Ericksson, N. Hamilton and B. Hankamer (2007). "Bilateral edge filter: photometrically weighted, discontinuity based edge detection." J Struct Biol **160**(1): 93-102.

Pantelic, R. S., W. Y. Fu, C. Schoenenberger and H. Stahlberg (2014). "Rendering graphene supports hydrophilic with non-covalent aromatic functionalization for transmission electron microscopy." Applied Physics Letters **104**(13): 134103.

Pantelic, R. S., R. Rothnagel, C. Y. Huang, D. Muller, D. Woolford, M. J. Landsberg, A. McDowall, B. Pailthorpe, P. R. Young, J. Banks, B. Hankamer and G. Ericksson (2006). "The discriminative bilateral filter: an enhanced denoising filter for electron microscopy data." J Struct Biol **155**(3): 395-408.

Park, W., G. Leibon, D. N. Rockmore and G. S. Chirikjian (2009). "Accurate image rotation using Hermite expansions." Image Processing, IEEE Transactions on **18**(9): 1988-2003.

PDB. "Protein data bank." Retrieved 06 June 2015, from http://www.rcsb.org/pdb/home/home.do.

Penczek, P. A. (2010). "Resolution measures in molecular electron microscopy." Methods Enzymol **482**: 73-100.

Petroutsos, D., A. M. Terauchi, A. Busch, I. Hirschmann, S. S. Merchant, G. Finazzi and M. Hippler (2009). "PGRL1 Participates in Iron-induced Remodeling of the Photosynthetic Apparatus and in Energy Metabolism in Chlamydomonas reinhardtii." Journal of Biological Chemistry **284**(47): 32770-32781.

Polle, J. E., S. D. Kanakagiri and A. Melis (2003). "tla1, a DNA insertional transformant of the green alga Chlamydomonas reinhardtii with a truncated light-harvesting chlorophyll antenna size." Planta **217**(1): 49-59.

Posten, C. and G. Schaub (2009). "Microalgae and terrestrial biomass as source for fuels--a process view." J Biotechnol **142**(1): 64-69.

Rafael C. Gonzalez, R. E. W. (2002). Digital Image Processing. Prentice-Hall.

Ramachandra, R., J. C. Bouwer, M. R. Mackey, E. Bushong, S. T. Peltier, N.-H. Xuong and M. H. Ellisman (2014). "Improving Signal to Noise in Labeled Biological Specimens Using Energy-Filtered TEM of Sections with a Drift Correction Strategy and a Direct Detection Device." Microscopy and Microanalysis: 1-9.

Ramakrishnan , V., T. Steitz and A. Yonath. (2009). "Noble Foundation." Retrieved 01/06/2015, from http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2009/.

Reichel, J., G. Menegaz, M. J. Nadenau and M. Kunt (2001). "Integer wavelet transform for embedded lossy to lossless image compression." IEEE Trans Image Process **10**(3): 383-392.

Renger, G. and T. Renger (2008). "Photosystem II: The machinery of photosynthetic water splitting." Photosynth Res **98**(1-3): 53-80.

Ress, D., M. L. Harlow, M. Schwarz, R. M. Marshall and U. J. McMahan (1999). "Automatic acquisition of fiducial markers and alignment of images in tilt series for electron tomography." J Electron Microsc (Tokyo) **48**(3): 277-287.

Roberts, L. G. (1963). Machine perception of three-dimensional soups, Massachusetts Institute of Technology.

Rochaix, J. D. (2001). "Assembly,function, and dynamics of the photosynthetic machinery in Chlamydomonas reinhardtii." Plant Physiol **127**(4): 1394-1398.

Roerdink, J. and A. Meijster (2001). "The Watershed Transform: Definations, Algorithms and Perallelization Strategies." Fundamenta Informaticae **41**: 187-228.

Roseman, A. M. (2003). "Particle finding in electron micrographs using a fast local correlation algorithm." Ultramicroscopy **94**(3-4): 225-236.

Rubinstein, J. L., J. E. Walker and R. Henderson (2003). "Structure of the mitochondrial ATP synthase by electron cryomicroscopy." EMBO J **22**(23): 6182-6192.

Ruskin, R. S., Z. Yu and N. Grigorieff (2013). "Quantitative characterization of electron detectors for transmission electron microscopy." J Struct Biol **184**(3): 385-393.

Sager, R. and G. E. Palade (1957). "Structure and development of the chloroplast in Chlamydomonas. I. The normal green cell." J Biophys Biochem Cytol **3**(3): 463-488.

Sandberg, K. (2007). "Methods for image segmentation in cellular tomography." Methods Cell Biol **79**: 769-798.

Sandberg, K. and M. Brega (2007). "Segmentation of thin structures in electron micrographs using orientation fields." J Struct Biol **157**(2): 403-415.

Saraste, M. (1999). "Oxidative phosphorylation at the fin de siecle." Science **283**(5407): 1488-1493.

Scheres, S. H. (2012). "RELION: implementation of a Bayesian approach to cryo-EM structure determination." J Struct Biol **180**(3): 519-530.

Scheres, S. H., R. Melero, M. Valle and J. M. Carazo (2009). "Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization." Structure **17**(12): 1563-1572.

Schermelleh, L., R. Heintzmann and H. Leonhardt (2010). "A guide to super-resolution fluorescence microscopy." J Cell Biol **190**(2): 165-175.

Schmidt, M., G. Gessner, M. Luff, I. Heiland, V. Wagner, M. Kaminski, S. Geimer, N. Eitzinger, T. Reissenweber, O. Voytsekh, M. Fiedler, M. Mittag and G. Kreimer (2006). "Proteomic analysis of the eyespot of Chlamydomonas reinhardtii provides novel insights into its components and tactic movements." Plant Cell **18**(8): 1908-1930.

Shaikh, T. R., H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith and J. Frank (2008). "SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs." Nat Protoc **3**(12): 1941-1974.

Shapiro, L., Stockman, G. (2002). Computer Vision, Prentice Hall.

Shapiro, L. G. and G. Linda (2002). "stockman, George C." Computer Vision, Prentice hall. ISBN 0-13-030796-3.

Shatsky, M., R. J. Hall, S. E. Brenner and R. M. Glaeser (2009). "A method for the alignment of heterogeneous macromolecules from electron microscopy." J Struct Biol **166**(1): 67-78.

Shi, Y. (2014). "A glimpse of structural biology through X-ray crystallography." Cell **159**(5): 995-1014.

Sijbers, J., P. Scheunders, M. Verhoye, A. van der Linden, D. van Dyck and E. Raman (1997). "Watershed-based segmentation of 3D MR data for volume quantization." Magn Reson Imaging **15**(6): 679-688.

Smil, V. (2006). Energy at the Crossroads. In OECD Global Science Forum - Conference on Scientific Challenges for Energy Research, OECD. Paris.

Sonka, M., V. Hlavac and R. Boyle (1999). Image Processing, Analysis, and Machine Vision., PWS Publishing, Pacific Grove, CA.

Sorzano, C. O., R. Marabini, J. Velazquez-Muriel, J. R. Bilbao-Castro, S. H. Scheres, J. M. Carazo and A. Pascual-Montano (2004). "XMIPP: a new generation of an open-source image processing package for electron microscopy." J Struct Biol **148**(2): 194-204.

Standfuss, J., A. C. Terwisscha van Scheltinga, M. Lamborghini and W. Kuhlbrandt (2005). "Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 A resolution." EMBO J **24**(5): 919-928.

Stark, H. and A. Chari (2016). "Sample preparation of biological macromolecular assemblies for the determination of high-resolution structures by cryo-electron microscopy." Microscopy (Oxf) **65**(1): 23-34.

Stephens, E., I. L. Ross and B. Hankamer (2013). "Expanding the microalgal industry--continuing controversy or compelling case?" Curr Opin Chem Biol **17**(3): 444-452.

Stephens, E., I. L. Ross, J. H. Mussgnug, L. D. Wagner, M. A. Borowitzka, C. Posten, O. Kruse and B. Hankamer (2010). "Future prospects of microalgal biofuel production systems." Trends Plant Sci **15**(10): 554-564.

Stoschek, A. and R. Hegerl (1997). "Denoising of electron tomographic reconstructions using multiscale transformations." J Struct Biol **120**(3): 257-265.

Stroebel, D., Y. Choquet, J. L. Popot and D. Picot (2003). "An atypical haem in the cytochrome b(6)f complex." Nature **426**(6965): 413-418.

Subramaniam, S. (2005). "Bridging the imaging gap: visualizing subcellular architecture with electron tomography." Curr Opin Microbiol **8**(3): 316-322.

Suga, M., F. Akita, K. Hirata, G. Ueno, H. Murakami, Y. Nakajima, T. Shimizu, K. Yamashita, M. Yamamoto, H. Ago and J. R. Shen (2015). "Native structure of photosystem II at 1.95 A resolution viewed by femtosecond X-ray pulses." Nature **517**(7532): 99-103.

Takahashi, H., M. Iwai, Y. Takahashi and J. Minagawa (2006). "Identification of the mobile light-harvesting complex II polypeptides for state transitions in Chlamydomonas reinhardtii." Proc Natl Acad Sci U S A **103**(2): 477-482.

Tang, G., L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees and S. J. Ludtke (2007). "EMAN2: an extensible image processing suite for electron microscopy." J Struct Biol **157**(1): 38-46.

Tani, K., C. P. Arthur, M. Tamakoshi, K. Yokoyama, K. Mitsuoka, Y. Fujiyoshi and C. Gerle (2013). "Visualization of two distinct states of disassembly in the bacterial V-ATPase from Thermus thermophilus." Microscopy (Oxf) **62**(4): 467-474.

Tokutsu, R., M. Iwai and J. Minagawa (2009). "CP29, a monomeric light-harvesting complex II protein, is essential for state transitions in Chlamydomonas reinhardtii." J Biol Chem **284**(12): 7777-7782.

Tolleter, D., B. Ghysels, J. Alric, D. Petroutsos, I. Tolstygina, D. Krawietz, T. Happe, P. Auroy, J. M. Adriano, A. Beyly, S. Cuine, J. Plet, I. M. Reiter, B. Genty, L. Cournac, M. Hippler and G. Peltier (2011). "Control of hydrogen photoproduction by the proton gradient generated by cyclic electron flow in Chlamydomonas reinhardtii." Plant Cell **23**(7): 2619-2630.

Tomasi, C. and R. Manduchi (1998). "Bilateral filtering for gray and color images." Sixth International Conference on . IEEE: 839-846.

Tomasi, C. and R.Manduchi (1998). Bilateral Filtering for Gray and Color Images. IEEE International Conference on Computer Vision, Bombay, India.

Tong, W.-S., C.-K. Tang, P. Mordohai and G. Medioni (2004). "First order augmentation to tensor voting for boundary inference and multiscale analysis in 3D." Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(5): 594-611.

Trepout, S., J. C. Taveau and O. Lambert (2010). "Structure determination of membrane protein by both cryo-electron tomography and single particle analysis." Methods Mol Biol **654**: 207-220.

Umena, Y., K. Kawakami, J. R. Shen and N. Kamiya (2011). "Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 A." Nature **473**(7345): 55-60.

van der Heide, P., X. P. Xu, B. J. Marsh, D. Hanein and N. Volkmann (2007). "Efficient automatic noise reduction of electron tomographic reconstructions based on iterative median filtering." J Struct Biol **158**(2): 196-204.

van Heel, M. (2013). "Finding trimeric HIV-1 envelope glycoproteins in random noise." Proc Natl Acad Sci U S A **110**(45): E4175-4177.

Villa, E., M. Schaffer, J. M. Plitzko and W. Baumeister (2013). "Opening windows into the cell: focused-ion-beam milling for cryo-electron tomography." Curr Opin Struct Biol **23**(5): 771-777.

Volkmann, N. (2002). "A novel three-dimensional variant of the watershed transform for segmentation of electron density maps." J Struct Biol **138**(1-2): 123-129.

Volkmann, N. (2004). "An approach to automated particle picking from electron micrographs based on reduced representation templates." J Struct Biol **145**(1-2): 152-156.

Volkmann, N. (2009). "Confidence intervals for fitting of atomic models into low-resolution densities." Acta Crystallogr D Biol Crystallogr **65**(Pt 7): 679-689.

Volkmann, N. (2010). "Methods for segmentation and interpretation of electron tomographic reconstructions." Methods Enzymol **483**: 31-46.

Volkmann, N. and D. Hanein (1999). "Quantitative fitting of atomic models into observed densities derived by electron microscopy." J Struct Biol **125**(2-3): 176-184.

Volkmann, N. and D. Hanein (2003). "Docking of atomic models into reconstructions from electron microscopy." Methods Enzymol **374**: 204-225.

Walters, R. G., D. G. Ibrahim, P. Horton and N. J. Kruger (2004). "A mutant of Arabidopsis lacking the triose-phosphate/phosphate translocator reveals metabolic regulation of starch breakdown in the light." Plant Physiol **135**(2): 891-906.

Weber, B., G. Greenan, S. Prohaska, D. Baum, H. C. Hege, T. Muller-Reichert, A. A. Hyman and J. M. Verbavatz (2012). "Automated tracing of microtubules in electron tomograms of plastic embedded samples of Caenorhabditis elegans embryos." J Struct Biol **178**(2): 129-138.

Weber, J. (2007). "ATP synthase--the structure of the stator stalk." Trends Biochem Sci **32**(2): 53-56.

Wei, Z., J. Wang, H. Nichol, S. Wiebe and D. Chapman (2012). "A median-Gaussian filtering framework for Moire pattern noise removal from X-ray microscopy image." Micron **43**(2-3): 170-176.

Whitelegge, J. P., H. Zhang, R. Aguilera, R. M. Taylor and W. A. Cramer (2002). "Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein: cytochrome b(6)f complex from spinach and the cyanobacterium Mastigocladus laminosus." Mol Cell Proteomics **1**(10): 816-827.

Wijffels, R. H. and M. J. Barbosa (2010). "An outlook on microalgal biofuels." Science **329**(5993): 796-799.

Wimberly, B. T., D. E. Brodersen, W. M. Clemons, Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch and V. Ramakrishnan (2000). "Structure of the 30S ribosomal subunit." Nature **407**(6802): 327-339.

Winkler, H. (2007). "3D reconstruction and processing of volumetric data in cryo-electron tomography." J Struct Biol **157**(1): 126-137.

Woolford, D., G. Ericksson, R. Rothnagel, D. Muller, M. J. Landsberg, R. S. Pantelic, A. McDowall, B. Pailthorpe, P. R. Young, B. Hankamer and J. Banks (2007). "SwarmPS: rapid, semi-automated single particle selection software." J Struct Biol **157**(1): 174-188.

Woolford, D., B. Hankamer and G. Ericksson (2007). "The Laplacian of Gaussian and arbitrary z-crossings approach applied to automated single particle reconstruction." J Struct Biol **159**(1): 122-134.

Woolford, D. S. A. (2007). Software advances and Applications in Cryo-EM. PhD Thesis, The University of Queensland, St. Lucia, Australia.

Yamashita, E., H. Zhang and W. A. Cramer (2007). "Structure of the cytochrome b6f complex: quinone analogue inhibitors as ligands of heme cn." J Mol Biol **370**(1): 39-52.

Yan, J., G. Kurisu and W. A. Cramer (2006). "Intraprotein transfer of the quinone analogue inhibitor 2,5-dibromo-3-methyl-6-isopropyl-p-benzoquinone in the cytochrome b6f complex." Proc Natl Acad Sci U S A **103**(1): 69-74.

Yu, Z. and C. Bajaj (2004). "Detecting circular and rectangular particles based on geometric feature detection in electron micrographs." J Struct Biol **145**(1-2): 168-180.

Zanetti, G., J. D. Riches, S. D. Fuller and J. A. Briggs (2009). "Contrast transfer function correction applied to cryo-electron tomography and sub-tomogram averaging." J Struct Biol **168**(2): 305-312.

Zhao, J., S. Benlekbir and J. L. Rubinstein (2015). "Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase." Nature **521**(7551): 241-245.

Zheng, S. Q., B. Keszthelyi, E. Branlund, J. M. Lyle, M. B. Braunfeld, J. W. Sedat and D. A. Agard (2007). "UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction." J Struct Biol **157**(1): 138-147.

Zhu, Y., B. Carragher, F. Mouche and C. S. Potter (2003). "Automatic particle detection through efficient Hough transforms." IEEE Trans Med Imaging **22**(9): 1053-1062.